Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020), pages 17–25
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Translating Knowledge Representations with Monolingual Word Embeddings: the Case of a Thesaurus on Corporate Non-Financial Reporting

Martín Quesada Zaragoza, Lianet Sepúlveda Torres, Jérôme Basdevant
Datamaran
Valencia, Spain
{martin, lianet, jerome}@datamaran.com

Abstract

A common method of structuring information extracted from textual data is using a knowledge model (e.g. a thesaurus) to organise the information semantically. Creating and managing a knowledge model is already a costly task in terms of human effort, not to mention making it multilingual. Multilingual knowledge modelling is a common problem for both transnational organisations and organisations providing text analytics that want to analyse information in more than one language. Many organisations tend to develop their language resources first in one language (often English). When it comes to analysing data sources in other languages, either a lot of effort has to be invested in recreating the same knowledge base in a different language or the data itself has to be translated into the language of the knowledge model. In this paper, we propose an unsupervised method to automatically induce a given thesaurus into another language using only comparable monolingual corpora. The aim of this proposal is to employ cross-lingual word embeddings to map the set of topics in an already-existing English thesaurus into Spanish. With this in mind, we describe different approaches to generate the Spanish thesaurus terms and offer an extrinsic evaluation by using the obtained thesaurus, which covers non-financial topics in a multi-label document classification task, and we compare the results across these approaches.

Keywords: thesaurus, cross-lingual word embeddings, bilingual lexicon induction, English, Spanish

1. Introduction

Corporate Social Responsibility (CSR) is a concept that aims to understand, categorise, monitor and regulate the actions of corporations regarding environmental, social, governance (ESG) and technological issues. Following Fontaine (Fontaine, 2013), one of the primary goals of CSR is to encourage corporations to engage in responsible behaviour when it comes to these issues (amongst others). CSR has become extremely relevant in the past decade. Customers, stakeholders and investors have increasingly begun to demand a robust integration of sustainable development practices into the wider business model - making sustainability a financially material issue. A growing number of policies and regulations, both voluntary and otherwise, have pushed companies towards public disclosure of non-financial (i.e. ESG) information in annual reports or stand-alone documents. The latest KPMG Survey of Corporate Responsibility Reporting (KPMG, 2017) indicates that 93% of the 250 largest companies by revenue (based on the Fortune Global 500 2016 index) have adopted non-financial reporting (NFR).

These corporate-derived disclosures are not the only factor to consider. The media, which informs about corporations and how they tackle ESG issues, is also a player when it comes to shaping discourse. Individuals, too, share on social media networks their views about organisations and sustainability. All these sources are relatively unstructured (i.e., they are only organised as natural language) textual data. As data scientists, we need to know what information we want to extract and how to organise it in a meaningful way if we want to gain insights and provide evidence for a data-driven decision making process. The sources that we are working with in this paper are courtesy of Datamaran, an ESG focused machine learning platform designed for material analysis of these issues. Datamaran already has an English-language thesaurus built to classify and structure data, which has been manually created and maintained by experts in sustainability matters. It covers over 100 topics and amounts to more than 6000 terms in an ongoing effort that has so far spanned over five years. However, analysing sources in English is only a part of the picture. If we really want to know what is happening in Spain or Latin America, we will need to be able to analyse texts in Spanish.

There are basically two options when it comes to annotating Spanish-language data:

1. to translate all the texts in Spanish into English and use our English thesaurus and pipeline to annotate the information in English, or
2. to create a thesaurus in Spanish so we can analyse texts in Spanish.

The first option seems at a glance to be the easiest and fastest solution. However, using a third-party translation API at scale is very expensive. Training your own Machine Translation (MT) model is not trivial, especially if you aim to translate from low-resource languages. The crux of the issue is to obtain appropriate machine-training data.

Manually creating a thesaurus in Spanish (or in any other language) would allow us to avoid the challenge of accurately translating large amounts of data. However, it would require finding experts in the field with command of the target language, or human translators with extensive ESG knowledge, and going through the process of terminology management and validation. This would be quite costly and slow. However, there is a valid option here if by substituting the thesaurus into our system, we can use the same automatic procedure to analyse text.

Our approach is based on using word embedding and cross-lingual mapping techniques in order to obtain seeds of terms for the Spanish thesaurus that correspond to the English thesaurus terms. Bearing this in mind, we evaluate the Artetxe et al. (2019) proposal, excluding the exposed unsupervised tuning procedure over the bilingual phrase table extracted from the cross-lingual mapping. Our primary objective is to obtain a mapping between the topics mentioned in English and Spanish. For that, we propose a set of heuristics to generate more terms in Spanish using the initial terms extracted from the cross-lingual mapping. The novelties of this proposal are: (i) an extrinsic evaluation of the Artetxe et al. (2019) approach on a multi-label document classification task and (ii) the creation of metrics to validate the quality of our results.

In Section 2. we provide an overview of the different approaches to solve the problem of analysing texts in a target language using a thesaurus in a different language. Next, we present the datasets used, we describe the experiments and the proposed heuristics in Section 3. The evaluation methodology is presented in Section 4.1. Later, we examine the results of the experiments and comment them in Section 5. Finally, we conclude with a summary of what we have learnt and remarks on future work in Section 6.

2. Related Work

Cross-lingual word embedding (CLE) techniques have raised and experienced rapid growth over the past few years, aided by the developments around neural word embeddings such as Word2vec (Mikolov et al., 2013a). Word embeddings are already a popular tool to induce bilingual lexicon, as continuous word embedding spaces exhibit similar structures across languages, even when considering distant language pairs (Mikolov et al., 2013b). Cross-lingual word embedding methods exploit these similarities to generate a common representation space that allows transferring information between two languages. Although early CLE techniques relied on partial parallel data, mapping-based CLE approaches only require seed translation dictionaries (Mikolov et al., 2013b; Faruqui and Dyer, 2014; Gardner et al., 2015) or no bilingual information at all (Artetxe et al., 2018a; Lample et al., 2018). The latter are especially effective in low-resource languages (Ruder et al., 2017) or specific domains. CLE facilitates a number of tasks that can benefit greatly from this unsupervised approach, one of which is bilingual lexicon induction (BLI).

Traditional BLI techniques extract word translations from monolingual corpora through a variety of monolingual distributional similarity metrics, such as orthographic, contextual and temporal similarity metrics to discover equivalent words in different languages (Haghighi et al., 2008; Knight, 2002). The popularity of CLE has encouraged research in applying both techniques together in order to induce a bilingual dictionary capable of obtaining successful results (Zhou et al., 2019; Søgaard et al., 2018a).

In this proposal we intend to use this latter BLI approach to generate a new thesaurus in a target language from a preexisting thesaurus in a different source language. We already possess an English thesaurus that groups a set of related lexical terms into topics or labels. For example, *acid rain, air contamination, air emission, air pollutant, air quality* are some of the terms grouped under the topic *Air Emissions*. Our main objective is to induce the English groups of terms that constitute each topic into Spanish, thus maintaining a topic alignment for both languages, but not necessarily a direct term equivalence.

Previous work on multilingual thesaurus alignment has already taken advantage of the structural correspondence between word embeddings in different languages and the semantic information that they offer to outperform alignment methods based on string similarity metrics (Gromann and Declerck, 2018). Our proposal further exploits these characteristics through the CLE mapping method VecMap (Artetxe et al., 2018a). VecMap is currently one of the most effective unsupervised CLE approaches, both in terms of BLI and cross-lingual document classification performance (Glavas et al., 2019). We chose this cross-lingual word embedding mapping method because of its performance and ease of use, as all of its code is publicly available and it works over the very common Word2vec toolkits. The method allows us to generate a new thesaurus in the Spanish language from a preexisting English thesaurus whilst avoiding any need for bilingual parallel data. To map the different labels or topics of our original thesaurus into another language, we translate each of their terms using a bilingual dictionary induced from a common representation space, according to the procedure described in Artetxe et al. (2018b). This cross-lingual space was previously generated from two monolingual word embeddings, following Artetxe et al. (2018a). We employ fastText (Bojanowski et al., 2016) to train the monolingual word embeddings. FastText is a Word2vec implementation that also captures sub-word information (Mikolov et al., 2013a). Unlike Word2vec, which trains the embedding considering the word as the smallest unit in a corpus, fastText learns word vectors at the character level of each word, which has a higher memory and time cost when compared to Word2vec. However, it is generally accepted that fastText performs better than Word2vec with out of vocabulary words, as it considers terms that do not appear in the training corpus.

Although more recent work introduces synthetic Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) models in order to generate sensible multi-word translation from a common representation space (Artetxe et al., 2019), we instead introduce some heuristics that perform adequately in the described thesaurus translation task and simplify the overall application of the method.

Unsupervised CLE tasks often prove to be hard to evaluate, especially when no manual standard exists for the downstream task at hand. In this proposal, we also offer a multi-label document classification evaluation method based on the annotation of a given parallel corpus that can be generalised to different thesauri.

3. Methodology

Our main objective is to use a cross-lingual word embedding and a thesaurus in a source language to generate a thesaurus in a target language without relying on parallel data. The source embeddings can be trained exclusively with data extracted from our ongoing analysis tasks, which is already available, easy to obtain and closely matches the characteristics of the information that will be analysed in the actual downstream use of the translated thesaurus. In this section we will describe the particular features of the knowledge base used in the application of this method, as well as characterise the bilingual lexicon induction techniques applied and our evaluation strategies. Finally, we propose an optimisation strategy for manual validation applied to the translated thesaurus.

3.1. Thesaurus

In order to structure the non-financial information, a thesaurus in English has been manually created by experts in sustainability matters. The thesaurus is a formal specification of concepts related to 100 NFR disclosure topics. For example, *air pollutant*, *air pollution*, *dust emission* are some of the concepts covering the topic *Air Emissions*. Our terms are both words or multi-word expressions and there are a significant quantity of noun phrases.

The thesaurus groups into topics more than 6000 terms in an ongoing effort that spans over five years. The terms of the thesaurus are expressed as lexical patterns to build a knowledge base of a matching algorithm responsible to automatically detect the mention of topics in different textual resources.

The patterns were created using the spaCy NLP library (Honnibal and Johnson, 2015). spaCy provides a rule-based matching feature that scans the text at token level according to some predefined patterns. These patterns are built considering some word-level features such as lemmatization, parts-of-speech, dependency parser and others. The matching algorithm compares the token attributes, specified by the pattern, with the attribute of the token in the input text to match or not the sequence. See below examples of the patterns we used.

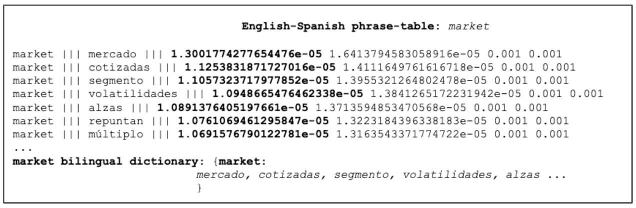
```
[
{"LOWER" : "dust"}, {"LOWER" : "emission"},
{"LOWER" : "diesel"}, { "LOWER" : "emissions"},
{"LOWER" : "air"}, {"LOWER" : "pollutant"}
]
```

In the above patterns, each element inside the square brackets represents one or more words that should appear consecutively. Each element inside the curly brackets represents a token. The *LOWER: diesel* means that we want to match a word whose lower form is *diesel*. For example, any of the following sequences will be annotated with the second pattern: *diesel emissions or DIESEL Emissions or Diesel emissions*.

Due to a lack of a Spanish thesaurus, we initially considered different alternatives to extract topics from Spanish texts: (1) maintaining a parallel thesaurus between source and target languages, which is a non-scalable process and required experts in target language; (2) using a Commercial Machine Translation System to translate the Spanish text into English. Although using a translation service seems a technically sound solution with adequate quality results, it is not financially feasible; or (3) training our own MT model, which requires too much effort and is also very costly. As a result, we moved on to BLI techniques to derive a Spanish thesaurus.

3.2. Building monolingual embeddings

To generate embeddings that can be used in the CLE method that we have selected for our translation purpose, we need two monolingual datasets: one in the source language in which our original thesaurus was built and another in the target language to which we want to migrate the said thesaurus. We apply lowercase and tokenization to both datasets, with which we then train two fastText embeddings with default hyperparameters and limiting the vocabulary to the 200,000 most frequent tokens as per Artetxe et al. (2019), although any Word2vec-based toolkit should suffice. The English and Spanish spaCy models were used to apply lowercase and to tokenize the datasets in both languages.



```
                    English-Spanish phrase-table: market

market ||| mercado ||| 1.3001774277654476e-05 1.6413794583058916e-05 0.001 0.001
market ||| cotizadas ||| 1.1253831871727016e-05 1.4111649761616718e-05 0.001 0.001
market ||| segmento ||| 1.1057323717977852e-05 1.3955321264802478e-05 0.001 0.001
market ||| volatilidades ||| 1.0948665476462338e-05 1.3841265172231942e-05 0.001 0.001
market ||| alzas ||| 1.0891376405197661e-05 1.3713594853470568e-05 0.001 0.001
market ||| repuntan ||| 1.0761069461295847e-05 1.3223184396338183e-05 0.001 0.001
market ||| múltiplo ||| 1.0691576790122781e-05 1.3163543371774722e-05 0.001 0.001
...
market bilingual dictionary: {market:
                          mercado, cotizadas, segmento, volatilidades, alzas ...
                          }
```

Figure 1: Fragment of market phrase-table and its candidate translations

3.3. CLE method

To obtain an inducted bilingual dictionary from monolingual data, we recreated the VecMap projection-based CLE method (Artetxe et al., 2018a) using the word embeddings mentioned in the previous section and mapped them into a shared space. We then extracted a source-to-target and target-to-source-phrase table using the technique described in Artetxe et al. (2018b).

A bilingual dictionary is directly induced from the source-to-target phrase-table by ranking the entries of a given term according to their corresponding likelihood in the phrase-table, thus transforming the quantitative ranking of the phrase-table into a bilingual dictionary with positional ranking. Figure 1 shows a fragment of the phrase-table obtained for the English term *market* and its Spanish candidate translations. Terms with higher likelihood will appear first in the entry for *market* in the induced bilingual dictionary dictionary.

This dictionary is used to translate the terms that make up our thesaurus. This approach maintains equivalence between source and target at the token level. However, many of the thesaurus terms are multi-word expressions. To cover this limitation and in order to build sensible combinations using the translated words, some heuristics are considered. As a result, token-level equivalence is often ignored.

3.4. Heuristics to generate terms in the target language

Using the cross-mapping embeddings we obtain a bilingual dictionary containing exclusively unigrams, which means that some techniques have to be applied in order to translate multi-word terms. In this section, we will outline several heuristic techniques that are applied to increase the coverage of the first bilingual dictionary. These heuristics use the phrase-table to generate new terms.

Literal translation Multi-word expressions are translated term by term, maintaining their original structure. The chosen translation for each word is the first-ranked one in the bilingual dictionary, or a special symbol if there is no possible translation for that term. This is the crudest possible form of translation using a bilingual unigram dictionary, and

Source language corpus	Target language corpus	Mean reciprocal rank
108,000 English news	84,000 Spanish news	0.093
220,000 English news	260,000 Spanish news	0.107

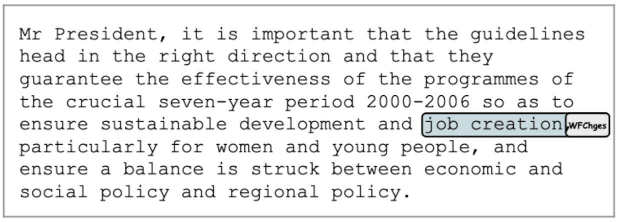
Table 1: Mean Reciprocal Rank that evaluates a bilingual dictionary against the full English to Spanish bilingual dictionary found in MUSE (Lample et al., 2018).

it serves as the baseline for all other heuristic approaches to building expressions. For example, for the English term **diesel emissions**, the literal translation that is obtained is ***diesel emisiones***, which can be represented as the following pattern:
[{"LOWER" : "emisiones"},{ "LOWER" : "diésel"}]

Permutations Expressions are first translated term by term, after which all of their possible permutations are added into the thesaurus. In languages that do not share a very similar grammatical structure, translating the expressions maintaining their original order may produce incorrect sentences. Moreover, this technique may help capture all possible variations in languages that present a flexible word order, such as Romance languages, Hungarian, etc. See below an example of the pattern obtained for the English term **diesel emissions** after obtaining its literal translation in Spanish and applying the permutation heuristic explained in this paragraph.
[{ "LOWER" : "diésel"},{"LOWER" : "emisiones"}]

Lemmatized terms with permutations All terms are translated in their original order, then lemmatized. Finally, like in the previous case, every possible permutation is considered. We lemmatize all terms in an attempt to reduce the variability that morphologically rich languages (that commonly also have a rather flexible word order) might bring, which is often a source of problems for unsupervised bilingual dictionary induction methods, as per Søgaard et al. (2018b). The following example shows the patterns generated using the current heuristic.
[
{"LEMMA" : "emisión"},{ "LEMMA" : "diésel"},
{ "LEMMA" : "diésel"},{"LEMMA" : "emisión"}
]

Lemmatized terms with permutations and wildcard inclusion We use the same setup as in the aforementioned approach, but adding a wildcard match before and after every word with the intent of boosting the coverage of the annotation. The longest possible match for each wildcard is selected, where the match can contain multiple tokens, and its sequence within the analysed text is no longer eligible for new matches. That is, we avoid overlap between different term matches. This logic might reduce the overall precision of the system, since overlap between the terms belonging to different labels is possible. We chose to operate in this manner to preserve the structure of our original thesaurus, as it does not present any overlaps between the terms of different labels. See below an example of one of the patterns generated adding the wildcard heuristic.
[
{"LEMMA" : "emisión"}, { "OP" : "*", "IS_ALPHA" : true}, {"LEMMA" : "diésel"}
]



```
Mr President, it is important that the guidelines
head in the right direction and that they
guarantee the effectiveness of the programmes of
the crucial seven-year period 2000-2006 so as to
ensure sustainable development and job creation WFChges
particularly for women and young people, and
ensure a balance is struck between economic and
social policy and regional policy.
```

Figure 2: Topic Workforce changes (WFChges) mentioned in an English Europarl sentence

4. Experimental settings

4.1. Data

The corpora necessary to build the initial monolingual word embeddings were generated using a preexisting collection of news articles from different online sources that are used in the Datamaran platform[1]. We chose to build these embeddings from news corpora because the English thesaurus that we intend to translate is used within Datamaran to analyse the content of online news, which would also be the purpose of this new translated Spanish thesaurus. Therefore, the domain of the corpora from which the monolingual word embeddings are built matches that of the text analysed in the downstream application of our system. The contents of the employed corpora are detailed below:

- Source language corpus, which contains 220,000 English news published during 2019, and more than 137,000,000 tokens.

- Target language corpus, composed by 260,000 Spanish news that appeared in online press during 2018-2019, containing around 118,000,000 tokens.

To validate the quality of the generated Spanish thesaurus we proposed a multi-label document classification task, that will be explained in Section 4.2.2. For that purpose, Version 7 of the English-Spanish Europarl corpus (Koehn, 2005) was used, as it contains a sufficient amount of terminology included in our particular thesaurus (datasets with very sparse annotation would not be very informative). Figure 2 shows an English sentence extracted form the Europarl corpus that mentioned the topic Workforce changes[2] (WFChges).

The Europarl corpus contains documents published on the European Parliament's official website, therefore it does not belong to the same domain as the corpus used to build the embeddings, which is a corpus of the news domain. This ensures that the performance obtained in the evaluation task

[1] https://www.datamaran.com/

[2] References to variation in number of people employed by an entity. Includes changes due to restructuring. E.g. reorganisations, turnover rates, outsourcing.

Translation Method	Phrase composition heuristic	Precision	Recall	KLD
VecMac (Artetxe et al., 2018a)	Literal translation	0.3871	0.2505	5.1149
VecMac (Artetxe et al., 2018a)	Permutations	0.5295	0.4590	1.6293
VecMac (Artetxe et al., 2018a)	Permutations and lemmatization	0.4236	0.5045	1.2235
VecMac (Artetxe et al., 2018a)	Permutations, lemmatization and wildcards	0.4580	0.6976	0.8027
Commercial Machine Translation System	None (the whole document is translated)	0.8209	0.8005	0.0233

Table 2: Multi-label document classification comparison over the parallel corpus Europarl for the English-Spanish pair. The embeddings used for the CLE approach (VecMap) were built from the corpora detailed in 4.1. Validation metrics at thesaurus level.

never surpasses what would be achieved when operating over a dataset that closely matched the information used to generate the embeddings, thus providing a pessimistic estimation for the effectiveness of the evaluated translated thesaurus. We find this property desirable, as it allows us to estimate the quality of the translation in the worst cases with a higher confidence level. Additionally, it can reveal faulty translations that could go undetected in a corpus of the same domain because of context similarities. For instance, the term "typhoons" is translated as "aviones" ("airplanes") in the bilingual dictionary generated with the techniques detailed in 3.3. using the aforementioned datasets. This could be because in news about typhoons it is usually mentioned that there will be delays or cancellations in commercial flights that operate in the affected region. However, airplanes are not necessarily mentioned next to typhoons in the Europarl corpus nearly as often, which means that when performing a multi-label document classification task it will be possible to appreciate that articles that only discuss the effects of reducing commercial flights in pandemics or passenger rights issues are getting labelled as if they were related to natural disasters.

4.2. Evaluation tasks

Even though CLE models are commonly evaluated considering only BLI tasks, their performance is heavily dependent on their actual application (Glavas et al., 2019), which highlights the need of using downstream performance evaluation tasks alongside BLI metrics.

4.2.1. BLI evaluation over a bilingual dictionary

A bilingual lexicon induction task is used to assess the quality of the bilingual dictionary generated as detailed in 3.3. This dictionary is compared against a ground-truth translation dictionary over the same language pair. The score of each term is obtained as the position of the first suggested translation for a term in the ground-truth bilingual dictionary within the list of possible translations for the same term of our induced bilingual dictionary, or zero if said translation is not included as an option in the generated dictionary. This scoring method is known as mean reciprocal ranking (MRR). MRR is equivalent to mean average precision (MAP) if only one valid translation per query is considered, in this case the top result. We chose this metric rather than the more common precision at rank k ($P@k$), which instead scores a term translation with one point if its position in the induced bilingual dictionary is equal or above k. This decision was made because MRR provides a more detailed evaluation, as it does not treat all models that rank the correct translation below k equally Glavas et al. (2019). In the evaluation, only terms from the ground-truth dictionary that had one or more of their possible translations appear in the induced translation dictionary were considered.

4.2.2. Multi-label document classification

Although BLI evaluation is a decent indicator our cross-lingual embedding quality along with the bilingual dictionary induce from it and can help the developer fine-tune this particular piece of the translation system, it does not directly correlate with the downstream performance of the system at hand, which in this particular use case corresponds to document classification. We propose a multi-label document classification task that directly matches the intended use of a translated thesaurus (where the classes of the task directly correspond to the topics of said thesaurus) and can be easily applied to other similar setups because of the simplicity of its logic.

The parallel bilingual corpus (i.e. the Europarl corpus referenced in section 4.1.) is considered, divided in different documents using an arbitrary window and one of the monolingual sections is annotated with the preexisting source language thesaurus (for our application, the source language in which the original thesaurus was written was English, so we would score the English section of the parallel corpus). This process will yield a score per document, which may have a different representation depending on the specific analysis criteria, be it mentions of a certain topic or the frequency of the terms with which it is related (i.e. combined incidences for all the terms that belong to a certain topic in a text). The source language thesaurus is then translated using a bilingual dictionary induced from two monolingual embeddings mapped into a common space, which are represented by two source-to-target and target-to-source cross-lingual embeddings, as seen in section 3.3.. This new thesaurus is used to annotate the section of the parallel corpus written in the target language (the Spanish section of the corpus in our case), thus obtaining a new list of document-score tuples. To get a better idea of how this parallel scoring would look in our case, we can see Figure 3, which shows two fragments of English and Spanish news extracted from https://elpais.com in which the topic Renewables alternatives[3] (Renew) was mentioned. Next to each highlighted term in Figure 3 we can see a label that indicates the topic

[3] References to energy from natural processes and/or non-traditional sources that are replenished on a human timescale. E.g. alternative energy sources, photovoltaic, biomass.

Topic code	Source frequency	Target frequency	Precision	Recall	Log-ratio	Priority
AEmiss	0.0112	0.0045	0.3269	0.2602	1.3067	0.1664
AltAccnt	0.0015	0.0012	0.5476	0.8214	0.2971	0.0006
AltFuel	0.0019	0.0103	0.0507	0.5625	-2.4477	0.2637
AntUse	0.0008	0.0046	0.1320	0.84	-2.4465	0.0528
AntiCorr	0.0314	0.0177	0.8649	0.8913	0.8272	0.8204
Biod	0.0075	0.0073	0.4701	0.8939	0.0398	0.0022
BrdComp	0	0.0002	0	0	-12.9659	0.0007
BuildAct	0	0.00005	0	0	-10.9665	0.00003
CChg	0.0234	0.0304	0.1633	0.4899	-0.3740	0.3465

Table 3: Multi-label document classification comparison over the parallel corpus Europarl for the English-Spanish pair. Validation metrics at topic level

to which it belongs (for instance "photovoltaic" is a term included in the topic Renew in our English thesaurus, and it appears in the English version of the article).

We now have two different scores per document, one obtained using the original thesaurus over the source language version of the article, and the other extracted with the induced thesaurus to rank the target language version of this same article. Based on the difference in label scoring per document we can obtain recall and precision at label (only scoring related to a specific topic is considered, i.e. hits from terms that belong to a specific topic) and thesaurus level (all topics are taken into consideration). We use micro averaging for both metrics, as the labels or topics of our thesaurus can present differences in the number of terms that they contain and how common those are. Furthermore, extracting the relative frequency of each label allows us to calculate the binary log of the ratio of relative frequencies (log-ratio) at label level and Kullback–Leibler divergence (KLD) (Kullback and Leibler, 1951) at thesaurus level.

Equation 1 is used to estimate the KLD, which quantifies how much a probability distribution diverges from another one in terms of the amount of additional information needed, where $P(x)$ and $Q(x)$ are the relative frequency of a category x in each corpora P or Q with respect to the categories in the same corpora. The higher the KLD value, the greater the divergence. If the value is 0, it means that both distributions are identical.

$$KLD(P||Q) = \sum P(x) log(\frac{P(x)}{Q(x)}) \tag{1}$$

Log-ratio (LR) is a metric commonly used for the task of keyword extraction, as it tells us how relevant the difference between two probability distributions is. It is estimated using Equation 2. We used the binary logarithm to better represent the dissimilitude pointed by the log-ratio measure.

$$LR(P||Q) = log(\frac{P(x)}{Q(x)}) \tag{2}$$

In our case, KLD compares the source and target language thesauri as a distribution of probabilities, where the relative frequency of each topic acts as the dependent variable, and the labels themselves are a qualitative factor (that is, the frequency is a topic or label-level metric, so its value will be different depending on the chosen topic). This comparison yields the expectation of the log difference between the probability of a topic in the original thesaurus distribution with the generated thesaurus distribution, which is the amount of information that is lost when we approximate the source language thesaurus with the target language one. The log-ratio is also given per topic and its estimation is based on the ratio of the relative frequency of a topic in the source and target corpora, providing a measure of how many times a topic is more frequent in one corpus compared with the other.

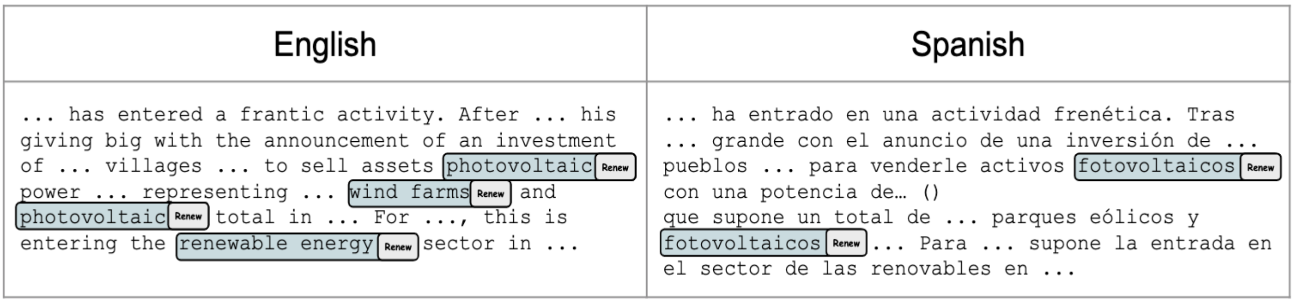
4.3. Optimising manual validation

No matter the performance of the technique in charge of translating a knowledge base, subsequent human validation will need to be applied in order to ensure the quality of the final product. This usually means that the thesaurus goes through a number of iterations before reaching its final state. However, knowledge bases can contain a tremendous volume of information, which complicates obtaining a complete human validation. With the objective of achieving an optimal partial validation, we establish a priority for each label or topic in our thesaurus and work over them in the resulting descending order. This priority metric guides the manual validation of the topic in the sense that topics with higher values should be the first to be reviewed manually, as they have a more significant impact over the quality of the translated thesaurus when compared with topics with lower priority. We achieve this by multiplying the absolute value of the log-likelihood ratio detailed in 4.2.2. with the source or target language frequency of the topic, depending on the sign of said log-ratio. For example, if the log-ratio is positive there are instances of the topic in the source language that are not being registered when analysing the text with the translated thesaurus in the target language. We multiply frequency of the topic in the original language with the absolute value of this positive log-ratio in order to get an idea of the negative impact of the translation of the aforementioned topic over the quality of our translated thesaurus.

$$Prio_i = |lr_i| * (of_i * (lr_i >= 0) + tf_i * (lr_i < 0)) \tag{3}$$

Where $Prio_i$ corresponds to the priority given to topic i, lr_i is the log-ratio obtained for topic i and of_i and tf_i are the original and target frequency for topic i respectively.

One of the side effects of using this formula is that topics that have close to no frequency at source and target will be classified as having low priority, even if their recall and



English	Spanish
... has entered a frantic activity. After ... his giving big with the announcement of an investment of ... villages ... to sell assets photovoltaic power ... representing ... wind farms and photovoltaic total in ... For ..., this is entering the renewable energy sector in ...	... ha entrado en una actividad frenética. Tras ... grande con el anuncio de una inversión de ... pueblos ... para venderle activos fotovoltaicos con una potencia de… () que supone un total de ... parques eólicos y fotovoltaicos ... Para ... supone la entrada en el sector de las renovables en ...

Figure 3: Multi-level classification for English and Spanish news using the thesauri in both languages

priority are zero or close to zero. We considered that, although recall and priority can be very low for a topic, if the source and target frequency are too low it becomes hard to assess the quality of the translation for the group of terms grouped under this topic, so human validation is not as useful as in other cases. Additionally, we can consider that uncommon terms will have a lower impact over the quality of the translated thesaurus even if their translations are not very good.

5. Results and discussion

Table 1 shows the MRR obtained from evaluating the bilingual dictionary generated from the base corpus described in the previous section (as well as a similar, smaller dataset) against the full English to Spanish bilingual dictionary provided in the MUSE specification (Lample et al., 2018). The bilingual dictionary evaluated is obtained according to the procedure described in 3.3. We compare a smaller corpus of online news against another dataset with a bigger volume that contains news from the same sources, the latter one being our main experimental corpus. Mean reciprocal rank for the generated bilingual dictionary does not always correlate directly with the actual downstream performance of the system, and some authors use it as a threshold of quality of the BLI procedure, like in Glavas et al. (2019), where 0.05 was established as a minimal value to consider a language pair translation run as acceptable. During our experiments, we have only considered MRR when measuring the impact of the size of the monolingual datasets with which our word embeddings are generated over the induced CLE-phrase table. It is displayed here to show how it can help developers evaluate certain pieces of the translation system individually (in this case our induced bilingual dictionary), and as a reference for future CLE-related tasks.

Conversely, our multi-label document classification evaluation (Table 2) yields much more informative results about the performance of both the source-to-target language alignment and the heuristic used to build terms from the induced unigram bilingual dictionary. As expected, literal translation returns low precision and recall scores, paired with a high KLD value, which indicates that most of the information contained in the original thesaurus is being lost. Part of the reason for this outcome can be attributed to the grammatical differences between Spanish and English, which are not properly accounted for when translating token by token.

Providing all possible permutations for each term has a notable impact for all metrics, but especially over the Kullback–Leibler divergence. Because KLD is a measure of information loss between two probability distributions (in this case modelled after the frequency of the topics in each thesaurus), we can infer that, although precision and recall are still relative low, this information loss is distributed more evenly across all the labels of the thesaurus. That is, the probability distributions that are modelled after topic annotation in the source and target language present a more similar shape.

Lemmatization seems to increase recall, which is expected, especially when working with a highly-inflectional language such as Spanish. However, it might introduce some noise, because it amplifies the coverage of all terms. This means that terms that were originally meaningful but that have been translated into common expressions will have a noticeable negative impact in the quality of the translated thesaurus. For instance, the term "unionized" that belongs to the topic "Union" is translated into Spanish as "trabajadores" (workers) in the bilingual English-Spanish dictionary obtained using the procedures detailed in 3.3. with the experimental settings mentioned in 4.1. "Trabajadores" is a much more common word that does not only appear in news articles concerning unionisation issues. This faulty translation already caused a loss in precision when using literal translation and permutation heuristics, but only in instances where the exact word appeared in the text. Moreover, the translation procedure depends on the actual contents of the used monolingual word embeddings, so it is possible that "trabajadores" often appears in a similar context to "unionized" and the precision is not affected excessively. However, after applying lemmatization, all possible forms of this term ("trabajadoras", "trabajador", "trabajadora") will produce a hit for the topic "Union".

Lastly, the addition of wildcards on top of the previous heuristics provides the best overall scores save for precision, which is still improved over using only permutations and lemmatization heuristics. The remarkable improvement of the recall is to be expected when applying this kind of "loose matching" (multiple tokens can appear in between the word that make up a multi-word term) over the Spanish language, which presents a flexible phrase structure. Even so, precision and KLD are still relatively far from the results obtained with the commercial machine translation system. In terms of precision, we have observed that our resulting bilingual dictionary has a tendency to place common terms as the most likely translation over more scarce expressions that may match the original term better. This phenomenon is likely related to the noise that results from the word embeddings cross-mapping procedure (Artetxe et al., 2018a). Further refinements in such processes and integration of the CLE-generated phrase-table into statistical or neural machine translation models may mitigate the issue, among other possibilities that we will briefly explore at the end of this section.

The results of our experiments show that the proposal does not perform equally for all the topics. This could be due to some topics being more or less specific, or due to factors that affect the number of occurrences of each topic in the training corpus. In Table 3 it is possible to see the frequency of each topic at source and target languages. For example, in the Europarl corpus we did not find mentions for any of the terms grouped under the topics BrdComp and BuildAct in the original English thesaurus, so the source frequency for both of these topics is 0. As a result, their precision and recall are zero independently of whether there are incidences for the same terms when translated with the method in our proposal (that is, their target frequency). However, this value for precision and recall does not imply that the translation of these two terms is necessarily bad. Instead, in cases where the source frequency for a topic is relatively low when compared to other topics, our confidence about the recall and precision values obtained will be lower. To reflect this, source and target frequency of the terms grouped under a topic contributes to the estimation of the priority of said topic, and the priority metric guides the manual validation of the topic in the sense that topics with higher values should be the first to review manually because they have worse results. For instance, the topic AntiCorr has a higher priority value, although it presents better precision and recall. In this case the priority metric is telling us that, even though this topic has been translated better than others, it appears very frequently in the analysed text, which means that it has a big impact over the quality of the translated thesaurus and should be reviewed before other topics. We can get to this conclusion because the priority is a function of the absolute value of the log-ratio and the frequency, which itself affects this calculation of the log-ratio the most. Consequentially, some topics have similar precision and recall values (i.e. Biod and AltAccnt), but the priority of one of them is lower (AltAccnt) because its terms are not very frequent. For cases where a topic has low values of precision and recall but its priority is still low, recommending additional terms for this topic could be useful.

Future improvements could include refining the phrase-table obtained from cross-lingual embeddings so as to obtain a better bilingual dictionary, as it has already been proposed in Artetxe et al. (2019), which also reduces the need for heuristics that build multi-word expressions. Term matching overlap can be tuned in order to maximise performance, although it would mean that the logic behind some of the terms of the original thesaurus might be compromised, which in some cases might be a better fit for the target language. It could also be of interest to evaluate terms that are found commonly as false positives according to their relevance (i.e. relying on tf–idf), discarding those that are too general by establishing a threshold and speeding up manual validation without losing meaningful terms.

6. Conclusion

In this work we offer a practical application of a bilingual lexicon induction (BLI) method based on cross lingual embeddings (CLE) (Artetxe et al., 2018a; Artetxe et al., 2018b) that allows us to induce a domain specific Spanish thesaurus from a preexisting English thesaurus used for multi-label document classification within Non-Financial Reporting. We include some possible heuristics that may help build sensible expressions from a unigram translation dictionary, which is itself induced from the aforementioned CLE procedure, and compare their performance against each other and a commercial machine translation system. To this end, we also offer some evaluation metrics that measure the performance of the proposed multi-label document classification task, along with a term prioritisation strategy for manual annotation. We hope that some of the strategies proposed here pave the way for an easier application of CLE-based BLI techniques, especially for tasks that rely on transferring information across multilingual knowledge representations, and help understand better the behaviour of these methods for similar use cases.

7. Acknowledgements

This paper was developed by a research team at Datamaran, Ltd. Datamaran is a company that develops AI-driven software for identifying and monitoring external risks and opportunities tied to ESG issues. We are thankful to our colleagues who provided expertise that greatly assisted the research.

8. Bibliographical References

Artetxe, M., Labaka, G., and Agirre, E. (2018a). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 789–798, Melbourne, Australia, July. Association for Computational Linguistics.

Artetxe, M., Labaka, G., and Agirre, E. (2018b). Unsupervised Statistical Machine Translation. Technical report.

Artetxe, M., Labaka, G., and Agirre, E. (2019). Bilingual Lexicon Induction through Unsupervised Machine Translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5002–5007, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information. 7.

Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 462–471, Gothenburg, Sweden, April. Association for Computational Linguistics.

Fontaine, M. (2013). Corporate social responsibility and sustainability: the new bottom line? *International Journal of Business and Social Science*, 4(4).

Gardner, M., Huang, K., Papalexakis, E., Fu, X., Talukdar, P., Faloutsos, C., Sidiropoulos, N., and Mitchell, T. (2015). Translation Invariant Word Embeddings. Technical report.

Glavas, G., Litschko, R., Ruder, S., and Vulic, I. (2019). How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. 2.

Gromann, D. and Declerck, T. (2018). Comparing pretrained multilingual word embeddings on an ontology alignment task. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May. European Language Resources Association (ELRA).

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. pages 771–779, 01.

Honnibal, M. and Johnson, M. (2015). An improved nonmonotonic transition system for dependency parsing. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1373–1378, Lisbon, Portugal, September. Association for Computational Linguistics.

Knight, K. (2002). Learning a translation lexicon from monolingual corpora. 05.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. Technical report.

KPMG. (2017). The KPMG Survey of Corporate Responsibility Reporting 2017. Technical report.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03.

Lample, G., Conneau, A., Ranzato, A., Denoyer, L., and Jégou, H. (2018). Word translation without parallel data. Technical report.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. 1.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting Similarities among Languages for Machine Translation. 9.

Ruder, S., Vulić, I., and Søgaard, A. (2017). A Survey Of Cross-lingual Word Embedding Models. 6.

Søgaard, A., Ruder, S., and Vulić, I. (2018a). On the limitations of unsupervised bilingual dictionary induction. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 778–788, Melbourne, Australia, jul. Association for Computational Linguistics.

Søgaard, A., Ruder, S., and Vulić, I. (2018b). On the Limitations of Unsupervised Bilingual Dictionary Induction. 5.

Zhou, C., Ma, X., Wang, D., and Neubig, G. (2019). Density matching for bilingual word embedding. In NAACL-HLT.