Can I find information about rare diseases in some other language?

Mikel Laburu	Alicia Pérez	Arantza Casillas	Iakes Goenaga	Maite Oronoz
IXA taldea	*IXA taldea*	*IXA taldea*	*IXA taldea*	*IXA taldea*
UPV-EHU	*UPV-EHU*	*UPV-EHU*	*UPV-EHU*	*UPV-EHU*
laburumikel@gmail.com	alicia.perez@ehu.eus	arantza.casillas@ehu.eus	iakes.goenaga@ehu.eus	maite.oronoz@ehu.eus

***Abstract*—Natural Language Processing (NLP) is a field that joins computer science and linguistics in an attempt to mimic, artificially, human language understanding. This paper applied NLP in the medical domain. The trigger that motivated this research was an expert reading an article about a rare disease who was interested in finding related documents. Being aware of the fact that language boundaries often limit, unnecessarily, the amount of information found, the goal of our work is to retrieve information without bounding to translation methods. Semantic similarity approaches offer a framework to represent related words and sentences in a dense space. In this work, we turned to cross-lingual dense spaces to represent bilingual documents in a shared dense space. Our approach helped to retrieve both intra- and cross-lingual documents just resting upon a few parallel documents to infer the optimal mapping from. From the experimental results we learned that an important issue is to keep aligned the mapping space and the cross-lingual search space. The cosine similarity outperforms both Euclidean and Manhattan distance. The results obtained in our preliminary experiments suggest that, although there is room for improvement, our approach performs satisfactorily achieving a P@10 of 71.72 searching English documents and returning Spanish related documents and 70.80 in the opposite direction.**

***Index Terms*—Clinical text mining, Cross-lingual information-retrieval, Natural language processing**

I. Introduction

Natural Language Processing (NLP) is an area of research and application that explores the aid of computers to understand and manipulate natural language text or speech.

The foundations of NLP lie in a number of disciplines, namely, computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, and psychology. Applications of NLP include a number of fields of study, such as machine translation, natural language text processing and summarization, user interfaces, speech recognition, artificial intelligence, and expert systems. Moreover, in the last years one of the most popular fields of study of NLP is multilingual and cross-language information retrieval (CLIR). This is because there is an increasing amount of full text material in various languages available through the Internet and other information suppliers. Therefore cross-language information retrieval has become an important new research area. It refers to an information retrieval task where the language of queries is other than that of the retrieved documents.

The need of CLIR systems in today's world is obvious. Moving from the global perspective to an individual level, CLIR is useful, for example, for users who are able to understand a foreign language but have difficulties in using it actively.

In this work the **aim** is to dive into medical abstracts from journals with the main goal of searching indistinctly related documents about rare diseases in English or in Spanish. From the methodological point of view, the key issue is that the focus is not on carrying out the requests by keywords or queries, instead, entire documents are taken. Furthermore, the result is given not only in the language of the source document but also in a different language. By contrast to other systems, our approach does not rest on machine translation and just requires few parallel documents to align the languages.

There is a further **motivation** on our work. It can be used, not only to seek information in journals but also in collections of document from other registers such as to retrieve bilingual electronic health records in places with co-official languages. Regardless of the language, patient records convey valuable information and convey relevant treatment that might help other practitioners on their decision making process. Thus, bridging linguistic gaps would help to have access to a wide range of information avoiding to discard valuable patient information in other languages. This is often the scenario in health systems comprising co-official languages with a dominant language and a minority language in their information processing system. As it is the case with the rare diseases, we feel that getting information in some language is more valuable than restricting the information to a given language.

II. Related work

The **trend** in CLIR is to translate either documents or queries [1], the goal is to produce a translation suitable for finding relevant documents written in a different language [2]. Research has concentrated on query translation, as it is computationally less expensive than document translation in terms of memory and processing capacity. This approach is more flexible than document translation and allows interaction with the user [2]. Within the query translation framework, basic approaches to CLIR are **machine translation** (MT), **corpus-based methods**, and **dictionary-based methods**:

- **MT** systems are often criticized by users for the quality of the translations, especially in the translation of an entire document with complex syntactic structures. Nevertheless, in recent years the quality of MT systems has improved using neural networks [3, 4, 5].
- In **corpus-based methods** queries are translated and expanded on the basis of multilingual terminology derived from comparable document collections or parallel corpora, these containing similar or identical documents in different languages [6, 7, 8].
- **Dictionary-based translation** [9, 10, 11] usually is an easier way to implement query translation if we compare it with the methods based on comparable documents or parallel corpora due to the fact that these ones are not always readily available. The main problems associated with dictionary-based CLIR are (1) untranslatable search keys due to the limitations of general dictionaries, (2) the processing of inflected words, (3) phrase identification and translation, and (4) lexical ambiguity in source and target languages. The category of untranslatable keys involves new compound words, special terms, and cross-lingual spelling variants, i.e., equivalent words in different languages which differ slightly in spelling, particularly proper names and loanwords.

McCarley [12] found that the efficiency relies on the translation direction more than on query or document translation. Apart from that, in some domains, for example in the medical domain, compiling enough corpus to develop MT systems is not always feasible. Moreover, get ready specialized dictionaries, a fundamental resource, is an extremely difficult task. As it is the case, corpora about rare diseases are extremely scarce. On the other hand, in recent years **semantic textual similarity** [13, 14] has become one of the most interesting research areas for NLP researchers. Semantic textual similarity (STS) measures the degree of semantic equivalence between two texts. Given two snippets of text, STS captures the notion that some texts are more similar than others, measuring their degree of semantic equivalence. Textual similarity can range from absence of relatedness to exact semantic equivalence. Accordingly, a grade is assigned as a similarity score that captures the notion of intermediate shades of similarity. Note that a given pair of texts may differ from some minor nuanced aspects of meaning to relatively important semantic differences, to sharing only some details, or to simply be unrelated in meaning. Recently, the Semeval 2017 Task 1 [14] has focused on semantic textual similarity for multilingual and cross-lingual pairs. The techniques explored in the task have many applications such as capturing a graded semantic relationship between two texts. Moreover, to promote improvement in other languages, the 2017 task draw attention to CLIR, among others, between English and Spanish. The best Spanish-English system made use of cross-language word embeddings [15]. Having analyzed the results and trends of the state of the art, we decided to create cross-lingual representations of documents to retrieve texts in both Spanish and English language. Our proposal focused on semantic similarity approaches without bounding to translation methods and just relying on a small dictionary. The proposed method is compatible with the problem of unavailability of specialized corpus.

III. Materials and methods

A. Corpus

EBCRD corpus [16] consists of set of medical abstracts in Spanish and English that cope with rare diseases. We have used a subset of these medical articles, to be precise the non-empty documents that were not repeated and had a counterpart in both languages, that is, unique translation pairs.

Quantitative details are given in Table I. The set was randomly divided into two disjoint sub-sets. One sub-set for the mapping step to project an embedded space into the other space (denoted as align) and the other sub-set to evaluate the information retrieval system (denoted as test). Table I reveals that, on average, documents in English and Spanish contain, respectively, 230 and 210 word-forms.

TABLE I: EBCRD corpus.

	English		**Spanish**	
	Align	**Test**	**Align**	**Test**
Documents	5,000	7,571	5,000	7,571
Words/doc	231.16	232.78	209.15	210.23

It is important to mention that it is difficult to find medical abstracts about rare diseases in any language. Therefore, the ability of finding the information cross-languages is crucial and very helpful for further documentation and research. The aim of our work is not to find translation pairs, conversely, we used these translation pairs to demonstrate that the system is able to find related information in different languages.

B. Creating cross-lingual dense spaces

It is well-known that bilingual features help in the disambiguation of clinical terms. An example of this kind of bilingual representation is the co-occurrence matrices of pairs of documents (each document being in a language) [16]. Bilingual embeddings were also used to enhance mono-lingual representations [17], using a method that learns bilingual embeddings based on multilingual knowledge bases.

Machine translation has also benefited from cross-lingual embedding mappings [18, 19], first training the monolingual word embeddings and then mapping them in the same space using bilingual dictionaries.

These are the steps followed to map source language embeddings into target language embeddings:

1) Generate document embeddings: Gensim software has been used to perform this task [20]. With this at the end of this process we get the dense representation of a given document e.g. $\mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$.
2) Normalize: Typically, we shall work with normalized vectors, $\|\mathbf{x}\| = 1$, meaning that the norm of all documents are in the surface of a hyper-sphere of radius

1. The motivation is to simplify the underlying vector algebra in the forthcoming mapping.

3) Map Source into Target language: there are different strategies to create cross-lingual dense spaces: Artetxe et al. [18] studied the possibility of mapping vector spaces using a very limited bilingual dictionary; Lample et al. [19] explored the possibility of learning to translate even without parallel data; Mikolov et al. [21] made one of the first approaches to map monolingual embeddings using bilingual data; in Zou et al. [22] a method to learn bilingual embeddings from a large unlabeled corpus was researched; Dinu et al. [23] proposed a method to correct the appearance of very frequent terms that can cause errors in the embeddings. The idea is to optimize a linear transformation in such a way that distances between translation pairs (usually words from a small dictionary) are minimized. The dictionaries do not necessarily cover all words. In fact, recent approaches focus on almost symbolic seed-dictionaries comprising just digits and punctuation marks or cognates [24]. In our case, the difference is that we are dealing with document embeddings and not word embeddings. Thus, we need a small set of aligned documents (what we refer to as *align* set), not necessarily aligned sentence by sentence.

The tool used to bring all the document embeddings to the same vector space was VecMap [25]. This requires a small dictionary to make the projection of one space into the other. To that end we used a small partition referred to as *Align* in Table I.

C. Similarity metrics

Given each document expressed as a point in a dense space (e.g. $\mathbf{x} \in \mathbb{R}^n$), we can search for close elements in this space. Formally, as expressed in (1), measured with distance d, the closest element to $\mathbf{x}$ is $\widehat{\mathbf{y}}$. Typically, we shall work with normalized vectors, $\|\mathbf{x}\| = 1$, meaning that the norm of all documents are in the surface of a hyper-sphere of radius 1.

$$\widehat{\mathbf{y}} = \arg\min_{\mathbf{x} \neq \mathbf{y}} d(\mathbf{x}, \mathbf{y}) \tag{1}$$

In order to define the closeness we can turn to well-known metrics (either similarity or di-similarity) such as Cosine similarity (d_C), Euclidean distance (d_E) or Manhattan distance (d_M), respectively, expressed in (2)-(4). Note that (2) is simplified due to the normalization boundary.

$$d_C(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i y_i \tag{2}$$

$$d_E(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{\frac{1}{2}} \tag{3}$$

$$d_M(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i| \tag{4}$$

D. Evaluation metrics for cross-lingual document retrieval

To assess cross-lingual document retrieval we turned to classical Information Retrieval metrics [26]. Let us help the reader interpret them in the context of this work:

- **Precision at k (P@k):** given documents in the language S, precision at k provides the relative number of times found the corresponding counterpart among the closest k documents in the language T. For each document in a source language, s_i, we find a set of closest documents in the target language, $\mathcal{T}^i = \{t_1^i, \cdots, t_k^i\}$ and next corroborate if the corresponding translation t_j is within $\mathcal{T}^i$ by means of $\delta(s_i, t_j)$ function that is 0 except if s_i and t_j are translation pairs, in which case is 1. Note that, against usual P@k from information retrieval, we consider that each document from S has just a single relevant result in T.

$$P@k = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{k} \delta(s_i, t_j^i) \tag{5}$$

- **Mean Reciprocal Rank (MRR):** while P@k informs about the presence of the target document within a set of k-closest documents, it does not take into account the position of the target document in the ranked list. Let us assume that given a document i (out of N) in the source language, we found the relevant document in the target language in position m_i, hence, the MRR of the set of N test documents is defined as (6).

$$MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{m_i} \tag{6}$$

- **Harmonic mean of the rank (Pos):** this value is the reciprocal value of the MRR. This gives us an idea of the position by which we expect to get the relevant document in the target language. It is defined as (7).

$$Pos = \frac{1}{MRR} \tag{7}$$

IV. Experimental results

We mapped the embeddings in two directions, from English to Spanish and the other way around. For completeness, the information retrieval was carried using, as source language, each of the two languages, besides, each set of documents, align and test, were explored separately. We employed three metrics to look for relatedness (cosine similarity, Euclidean distance, Manhattan distance). In this context, the ability of the mapped embeddings technique to find information **cross-languages**, was assessed in Table II.

From these results we learned the following lessons:

1) The best text-similarity measure was the cosine similarity. This relatedness approach (cosine similarity) optimized all the assessment criteria (P@k, MRR, Pos) in all the data sets (align, test) for both languages (Spanish, English).

2) The mapping direction matters: better results are achieved searching from language S to T whenever the mapping is carried out in the same direction.
3) Slightly better results are achieved on the align set used to map the embedding spaces than with the independent test set, however, the differences are not significative. This corroborates that the method is able to learn relatedness in aligned sets and extend without significant increment of errors into non-aligned sets.

With these insights we decided to extended the results of P@k varying the k using the cosine distance to search the documents having mapped the embedding spaces in the same direction as the search as shown in Fig 1. This figure revealed that both searching directions (English to Spanish and the reverse) achieved similar results, both curves are almost overlapped for each set (align and test). This implies that the system is equally robust regardless the search direction.

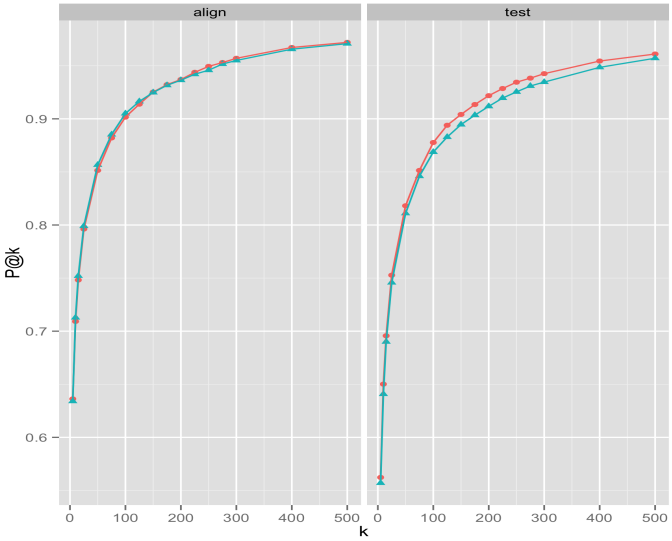

Fig. 1: P@k (ordinate) varying k (abscissa) with equal embedding mapping and search direction, both English to Spanish (●) or Spanish to English (▲). Cosine similarity was used to retrieve similar information in the other language.

So far, we assessed the ability of the system to retrieve information related to a source language in a target language, that is, cross languages. Next we turned to **intra-language** information retrieval looking for a document in the source language within the set of documents in the source language. We verified that, regardless of the similarity metric used (cosine similarity, Manhattan distance, Euclidean distance) the first document in the rank was 100% of the times the document itself. That is, the identity property associated to any metric or distance function is satisfied.

In an attempt to enable the reader focus on tangible results, we would like to show three examples (abstracts) and next, through Table III, show the cosine similarity attained in the cross-lingual dense space for each pair of documents varying both the mapping direction and the search direction. The **examples** are as follows:

A:

"Achalasia is an incurable primary esophageal motor disorder of unknown aetiology. The intent of any treatment is to weaken the lower esophageal sphincter. Established treatments for endoscopic management are endoscopic application of botulism toxin and pneumatic dilation, along with other treatments under development such as POEM (per-oral endoscopic myotomy). The first 2 are very effective in weakening lower esophageal sphincter pressure, but their efficacy and duration vary greatly. There is a recurrence of symptoms of 50% at 6 months and almost 100% in one year for botulism toxin, while with pneumatic dilation only 60% of patients are asymptomatic at 5 years, so the need for multiple pneumatic dilation is the rule in $>$ 90% of patients. The best positive predictor of lasting symptomatic response is lower esophageal sphincter pressure $<$ 15 mmHg after performing any procedure. The POEM technique is promising and still evolving, initially offering results similar to the Heller myotomy; however, we are waiting for greater experience with more patients and for long-term results."

B:

"Achalasia is an infrequent esophageal disease that severely impairs the quality of life of affected individuals. The etiology of this entity is not well defined and its main clinical features are dysphagia and regurgitation. The treatment of achalasia continues to be palliative and is aimed at providing functional and symptomatic relief through opening of the lower esophageal sphincter. The present article describes and evaluates the medical and surgical treatments most commonly used in clinical practice after the introduction of minimally invasive surgery, which has represented an important addition to the therapeutic alternatives. Currently, the most appropriate initial option is laparoscopic surgery, while pneumatic dilatation and botulinum toxin injection should be reserved for selected patients."

C:

"Background and objective: To estimate the prevalence of obesity and overweight in Canary adolescents, evaluating its association with breakfast intake and physical activity. Subjects and method: Cross-sectional study of a representative sample of children aged 12-14 years living in the island of Gran Canaria. They were weighed and measured and the prevalence of overweight and obesity was determined according to the 85th and 97th percentiles of the Spanish body mass index tables. Breakfast and physical activity characteristics were also studied using questionnaires. Results: The overall prevalence of obesity and overweight was 26.1%, higher in females (29.5%) than in males (22.8%). Obesity affects 14.8% of all teenagers (17.6% of girls and 12.0% of boys). Highest overweight and obesity levels affect those aged 12 years, decreasing progressively with age. Those boys who have a more complete breakfast have lower prevalence rates. There was no association between obesity and overweight with physical activity, as measured by the number of hours devoted to watching television or playing videogames as opposed to hours devoted to sport. Conclusions: The prevalence of

TABLE II: Cross-lingual information retrieval results. Given a document in a source language we look for similar documents in a target language. The document-embedding mapping-direction was from source to target and conversely. Two sets were seek, the one used to create the mappings (Align) and the independent test (Test). Three similarity metrics were considered: Cosine (Cos), Euclidean (Euc) and Manhattan (Man). Three metrics were used to assess the system: P@k with k=10, MRR and Pos.

Mapping		**Search**			**P@10**			**MRR** (10^{-3})			**Pos**		
Source	**Target**	**Source**	**Target**	**Set**	**Cos**	**Euc**	**Man**	**Cos**	**Euc**	**Man**	**Cos**	**Euc**	**Man**
English	Spanish	English	Spanish	Align	71.72	69.38	65.56	16.93	13.11	10.66	59.04	76.24	93.78
				Test	66.10	63.37	58.12	111.84	9.09	6.78	84.39	109.89	147.36
		Spanish	English	Align	51.42	18.44	5.44	9.11	1.36	0.58	109.71	734.63	1,704.79
				Test	46.11	15.45	4.57	6.86	1.22	0.53	145.73	818.22	1,884.35
Spanish	English	English	Spanish	Align	55.26	23.24	8.84	9.79	1.41	0.58	102.05	705.21	1,714.38
				Test	50.17	20.12	6.97	7.74	1.26	0.53	129.17	792.03	1,874.31
		Spanish	English	Align	70.80	69.68	65.55	15.73	14.56	10.82	63.56	68.65	92.37
				Test	64.81	63.66	59.39	11.34	11.12	8.26	88.18	89.89	120.99

overweight and obesity is high, especially in girls. We observed an inverse relationship between breakfast and its quality and obesity."

Table III shows the intra-lingual cosine similarity between these three documents (whenever the source and target language is the same) and also cross-lingual (when the source and target languages differ). To this end, the embedding mapping direction and the search direction were aligned. Color-scale refers to the cosine similarity as a heat-map where total similarity (cosine equal to 1.00) is highlighted in orange and absence of similarity (negative cosine) in blue. For example, documents identified by B and A are semantically related, both focus on achalasia and their treatments. From the cosine similarity between them (see Table III) we can draw the same conclusion. Moreover, Table III shows that both documents are closely related regardless the search direction, which agrees with the human notion of semantical relatedness that goes beyond languages. On the other hand, documents B and C deal with barely related topics (while the former deals with achalasia, the later focuses on teenagers' obesity) and so indicate the negative similarity in Table III associated to this pair of documents. For the sake of curiosity, we would like to mention that both A and C belong to the align set while B belongs to the test set.

From Fig. 1 we learned that, on average, the precision achieved is independent from the search direction. Nevertheless, paying attention at individual documents, as in Table III, we found that the distance from document A to B with the search direction S to T is not always the same as with the reverse search direction, i.e. $d(A_S, B_T) \neq d(A_T, B_S)$. Note that, while the cosine similarity is symmetric, this property seems to be distorted, possibly due to the space mapping operation. It seems that this distortion is homoscedastic for the overall impact obtaining almost identical results in both searching directions (Fig. 1). Again, locally, it is easy to guess what we corroborated globally, that is, the importance of carrying out the search in the mapping direction. For example, with the space mapped from English into Spanish (Table IIIa), the document A in English has a similarity of 0.58 with the document B in Spanish; by contrast, with

TABLE III: Intra-lingual and cross-lingual cosine similarity between three documents (A, B and C). A document in the source language was compared to another in the target language (for all English and Spanish language-pair combinations) varying search directions. Each sub-tables shows the results achieved using a different mapping direction.

			Target					
			English			Spanish		
			A	B	C	A	B	C
Source	English	A	1.00	0.51	-0.29	0.64	0.58	-0.32
		B	0.51	1.00	-0.35	0.58	0.73	-0.57
		C	-0.29	-0.35	1.00	-0.32	-0.57	0.65
	Spanish	A	0.64	0.63	-0.33	1.00	0.67	-0.53
		B	0.63	0.73	-0.42	0.67	1.00	-0.73
		C	-0.33	-0.42	0.65	-0.53	-0.73	1.00

(a) Mapping: source English, target Spanish

			Target					
			English			Spanish		
			A	B	C	A	B	C
Source	English	A	1.00	0.53	-0.46	0.61	0.48	-0.35
		B	0.53	1.00	-0.37	0.48	0.68	-0.35
		C	-0.46	-0.37	1.00	-0.35	-0.35	0.63
	Spanish	A	0.61	0.48	-0.41	1.00	0.73	-0.54
		B	0.48	0.68	-0.53	0.73	1.00	-0.63
		C	-0.41	-0.53	0.63	-0.54	-0.63	1.00

(b) Mapping: source Spanish, target English

the space mapped from Spanish to English (Table IIIb), the document A in English has a similarity of 0.48 with the document B in Spanish, smaller than before: $0.58 = d_C(E_A, S_B)|_{Map(E,S)} > d_C(E_A, S_B)|_{Map(S,E)} = 0.48$. Nevertheless, locally, this property is not always filled, for example, $0.61 = d_C(S_A, E_B)|_{Map(S,E)} \not> d_C(S_A, E_B)|_{Map(E,S)} = 0.64$

This work shows an ongoing and promising field of research. Nevertheless there are many open research questions. We restricted to document embeddings of dimension $n = 300$ created with doc2vec [27] within Gensim library [20]. We feel that it is well-worthy fine-tuning the embeddings with other embedding generation techniques (such as FastText [28], Elmo [29] etc.) and other dimensions. The popular cosine similarity outperformed the other two distances used (Euclidean and

Manhattan), however, there are many more metrics well suited to normalized spaces that remained out of the scope of this article, not to mention other space transformations (logarithmic or hyperbolic transformations) that might leverage proximity assessment criteria. Moreover, we feel curious about the impact of the size of the bilingual seed (the align set) in the quality of the semantic relatedness.

V. Concluding remarks and future work

The aim of this work was to explore cross-lingual information retrieval. The motivation is that the language should not imply a shortcoming for information systems. Recovering critical and, often, rare information is important regardless the language, particularly, within the clinical domain. Traditional information retrieval systems are based on queries or keywords. By contrast, for rare information and infrequent information retrieval the user might be hesitant with the selection of accurate keywords. We would like to make a step ahead and help the user navigate in a natural way with a fuzzy search in which the entire document serves as search goal. The final goal is to find similar documents to a given one without language limitations.

To bridge the gap between languages, we resorted to cross-lingual dense spaces. These spaces were previously used in machine translation to project bilingual words (discrete pairs) in continuous spaces (as numeric vectors). In this work we projected entire documents in a single space. Interestingly enough, similar properties to those found for words were applied to entire documents. That is, we found that close documents in the cross-lingual dense space were semantically related. Besides, intra-lingual relations are kept as well as cross-lingual relations.

The experimental results provided the following insights: 1) the highest semantic similarity was obtained mapping the cross-lingual spaces in the same direction as the search; 2) cosine similarity is more appropriate than either Euclidean or Manhattan distance in this context; 3) the method barely deteriorates from the align set to the test set, meaning that a small aligned seed gives the chance to extrapolate the location of documents in the dense space.

This is a preliminary work that opened a promising framework within clinical information retrieval. This work brings interesting open questions ahead that are worthy exploring, such as different document embedding strategies and representations ($\mathbf{x}$), alternative similarity metrics (d), and the sensitivity of the approach with respect to the size of the aligned seed. Our intuition is that cross-lingual dense spaces downgrade the quality of the intra-language information retrieval at the expense of enabling cross-lingual retrieval. In the near future we plan to delve into this research question and assess the factors that affect most in the space mapping stage.

Acknowledgment

This work was partially funded by the Spanish Ministry of Science and Innovation (PROSAMED: TIN2016-77820-C3-1-R and TADEEP: TIN2015-70214-P) and the Basque Government (DETEAMI: 2014111003, BERBAOLA: KK-2017/00043).

References

[1] J. Chin, M. Heymans, A. Kojoukhov, J. Lin, and H. Tan, "Cross-language information retrieval," Aug. 5 2014, uS Patent 8,799,307.

[2] J.-Y. Nie, *Cross-Language Information Retrieval*. Morgan & Claypool Publishers series, 2010.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[4] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[5] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[6] P. Sheridan and J. P. Ballerini, "Experiments in multilingual information retrieval using the spider system," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1996, pp. 58–65.

[7] K. Yamabana, "A language conversion front-end for cross-linguistic information retrieval," in *Proceedings of SIGIR Workshop on Cross-Linguistic Information Retrieval, Zurich, Switzerland, 1996*, 1996.

[8] M. Davis, "New experiments in cross-language text retrieval at nmsus computing research lab. november 1996. approach of the new mexico state university," *URL: http://crl. nmsu. edu*.

[9] W. Kraaij and R. Pohlmann, "Different approaches to cross language information retrieval," *LANGUAGE AND COMPUTERS*, vol. 37, pp. 97–110, 2001.

[10] M. Adriani, "Dictionary-based clir for the clef multilingual track." in *CLEF (Working Notes)*, 2000.

[11] L. Ballesteros and B. Croft, "Dictionary methods for cross-lingual information retrieval," in *International Conference on Database and Expert Systems Applications*. Springer, 1996, pp. 791–801.

[12] J. S. McCarley, "Should we translate the documents or the queries in cross-language information retrieval?" in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, 1999, pp. 208–214.

[13] E. Agirre, M. Diab, D. Cer, and A. Gonzalez-Agirre, "Semeval-2012 task 6: A pilot on semantic textual similarity," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, ser. SemEval '12.

Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 385–393. [Online]. Available: http://dl.acm.org/citation.cfm?id=2387636.2387697

[14] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation," *arXiv preprint arXiv:1708.00055*, 2017.

[15] J. Ferrero, F. Agnes, L. Besacier, and D. Schwab, "Compilig at semeval-2017 task 1: Cross-language plagiarism detection methods for semantic textual similarity," *arXiv preprint arXiv:1704.01346*, 2017.

[16] A. Duque, J. Martinez-Romo, and L. Araujo, "Can multilinguality improve biomedical word sense disambiguation?" *Journal of biomedical informatics*, vol. 64, pp. 320–332, 2016.

[17] J. Goikoetxea, A. Soroa, and E. Agirre, "Bilingual embeddings with random walks over multilingual wordnets," *Knowledge-Based Systems*, vol. 150, pp. 218–230, 2018.

[18] M. Artetxe, G. Labaka, and E. Agirre, "Learning bilingual word embeddings with (almost) no bilingual data," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2017, pp. 451–462.

[19] G. Lample, L. Denoyer, and M. Ranzato, "Unsupervised machine translation using monolingual corpora only," *arXiv preprint arXiv:1711.00043*, 2017.

[20] "Gensim: Topic modelling for humans," https://radimrehurek.com/gensim/.

[21] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *arXiv preprint arXiv:1309.4168*, 2013.

[22] W. Y. Zou, R. Socher, D. Cer, and C. D. Manning, "Bilingual word embeddings for phrase-based machine translation," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1393–1398.

[23] G. Dinu, A. Lazaridou, and M. Baroni, "Improving zero-shot learning by mitigating the hubness problem," *arXiv preprint arXiv:1412.6568*, 2014.

[24] M. Artetxe, G. Labaka, and E. Agirre, "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)*, 2018.

[25] M. Artetxe, G. Labaka, and E. Agirre, "Learning principled bilingual mappings of word embeddings while preserving monolingual invariance," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2289–2294.

[26] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press, 2008, vol. 39.

[27] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*, 2014, pp. 1188–1196.

[28] "Fasttext: Library for efficient text classification and representation learning," https://fasttext.cc/.

[29] "Elmo," https://allennlp.org/elmo.