Semantic Recommendation System for Bilingual Corpus of Academic Papers

Anna Safaryan[1], Petr Filchenkov[1(✉)], Weijia Yan[1], Andrey Kutuzov[2], and Irina Nikishina[3]

[1] National Research University Higher School of Economics, Moscow, Russia
anna.safaryan-813@yandex.ru, psfilchenkov@edu.hse.ru, renatayanweijia@gmail.com
[2] University of Oslo, Oslo, Norway
andreku@ifi.uio.no
[3] Skolkovo Institute of Science and Technology (Skoltech), Moscow, Russia
Irina.Nikishina@skoltech.ru

Abstract. We tested four methods of making document representations cross-lingual for the task of semantic search for the similar papers based on the corpus of papers from three Russian conferences on NLP: Dialogue, AIST and AINL. The pipeline consisted of three stages: preprocessing, word-by-word vectorisation using models obtained with various methods to map vectors from two independent vector spaces to a common one, and search for the most similar papers based on the cosine similarity of text vectors. The four methods used can be grouped into two approaches: 1) aligning two pretrained monolingual word embedding models with a bilingual dictionary on our own (for example, with the VecMap algorithm) and 2) using pre-aligned cross-lingual word embedding models (MUSE). To find out, which approach brings more benefit to the task, we conducted a manual evaluation of the results and calculated the average precision of recommendations for all the methods mentioned above. MUSE turned out to have the highest search relevance, but the other methods produced more recommendations in a language other than the one of the target paper.

Keywords: Semantic similarity · Semantic search · Scientific literature search · Document representations · Cross-lingual representations

1 Introduction

Only a couple of decades ago, search engines were mostly based on literal occurrences of the query words in the documents. Nowadays, we are witnessing the development of another approach: semantic search, in general terms, is a search with meaning, that can be extracted from a query, some data or an ontology. It enables a search engine user to find relevant pieces of information irrespective of the mentioned formal criteria, i.e. the search results may not contain the query words. The exact definition of the term "semantic search" is quite ambiguous, and we refer the reader to the survey [7] for more details.

W. M. P. van der Aalst et al. (Eds.): AIST 2020, CCIS 1357, pp. 22–36, 2021.
https://doi.org/10.1007/978-3-030-71214-3_3

Performing semantic search in a multilingual corpus of documents is a challenging task. Even within one language there is a huge variety of grammatical, cultural and pragmatic aspects which make the meaning ambiguous. Transferring all these aspects to another language seems extremely confusing. A good start to test multilingual semantic search is therefore to perform it in a corpus of particular domain. Narrowing the corpus to two languages could also facilitate the task. That is why this research deals with the bilingual corpus of academic papers. But its results can be extended to other corpora and languages.

There is a vast number of academic corpora such as a corpus of Wikipedia articles on science and technology [16], a corpus of academic journal papers [9] or any academic sub-corpus of national corpora (BNC[1], RNC[2], etc.). We test several approaches on a similar bilingual corpus of academic papers associated with the RusNLP project [18][3]. RusNLP is a search engine for academic papers presented in Russian NLP conferences: Dialogue, AIST and AINL. This project currently operates with papers in English, although there are still some papers written in Russian.

The research question of the paper stems from the practical issue of using cross-lingual word embedding models for semantic recommendation system on the corpus of academic papers. We have tested the following models: result of a simple word-by-word translation; result of projecting vectors from a Russian model vector space to an English one [14]; Multilingual Unsupervised and Supervised Embeddings (MUSE) [10]; a supervised model based on VecMap framework [2] (more details are given in the Sect. 3.2). Since only one of them was off-the-shelf, we would like to discover, whether it is enough to take pre-aligned word embeddings to create a decent recommendation system or it is better to take independent pretrained monolingual models and try to map them into common vector space in different ways. To the best of our knowledge, there is no other research comparing various approaches to cross-lingual document recommendation on a bilingual corpus of academic papers. Additionally, the most promising approach can be implemented for the RusNLP web service.

The paper is organised as follows: a brief overview of the existing papers on cross-lingual retrieval methods is given in Sect. 2; a more detailed description of the dataset and methods implementation is provided in Sect. 3; the evaluation of implemented cross-lingual search methods is outlined in Sect. 4.

2 Related Work

At the moment, there are many described attempts to implement a cross-lingual retrieval on academic papers [6,22]. However, none of them employs cross-lingual word embedding models—a tool that reflects the similarity of words in different languages. Word embedding models infer distributed representations of words in

[1] https://www.english-corpora.org/bnc/.
[2] http://www.ruscorpora.ru.
[3] https://nlp.rusvectores.org.

a low dimensional continuous space. Such models may be based on neural networks predicting the context words or low-rank approximations of word-context matrices [19]. At the core of word embeddings lies the distributional hypothesis that can be formulated as follows: the meanings of words used in similar contexts tends to be the same. We can notice that monolingual distributional models are widely employed in various NLP domains [26]. Cross-lingual models are closely connected with the monolingual ones and have a solid number of architectures to be tested. A confusion may occur as they are often based on a combination of different approaches, and the number of publications, meanwhile, grows. [21] is an attempt to provide a thorough survey of cross-lingual word representations. The authors systematise the main types of both mono- and cross-lingual word embedding models, describe their internal structure and working principles.

Supervised learning of cross-lingual embeddings often involves aligning monolingual models: a transformation of vector spaces on the basis of a multilingual dictionary that allow us to make corresponding word embeddings from different language close to each other. One of the first methods was linear transformation, proposed in [14]. Among state-of-the-art alignment methods there are such algorithms as Multilingual Unsupervised and Supervised Embeddings (MUSE) [10], Word Embedding Mapping [2], and Multilingual BERT [20]. A supervised learning may have some restrictions but still is in a focus of attention [17].

Weakly supervised cross-lingual embeddings are another possible method that can be used to deal with the lack of parallel data. A smaller dictionary is the main difference that distinguishes this approach from supervised learning. The paper [1] proposes a bilingual word embedding model with a 25-word dictionary. In [28], we can observe that even as few as 10 seed words are enough to make an alignment between embeddings of two different languages to perform multilingual POS tagging.

Many recent studies propose unsupervised learning of the cross-lingual word embeddings [12,25]. The key advantage of such models is their capability to be trained completely or almost without parallel data. Low-resource languages sometimes leave no choice but to use unsupervised models. In [3], the authors argue that strict unsupervised training without any parallel data is rather impractical. Nevertheless, they acknowledge theoretical scientific value of further research in this direction.

In the main experiment to which this paper is devoted, only word-level supervised methods to align monolingual embeddings were used in addition to the baseline without any cross-lingual embeddings. Their more detailed description is provided below:

- **Linear transformation** of vectors from the one vector space to another. The least squares method is used to calculate a matrix that, when applied to a word vector from the source language, transforms it into a vector which is as close as possible to the word vector of its translation in the target language, i.e., into the target vector space [14].
- **Multilingual Unsupervised and Supervised Embeddings** (**MUSE**) uses singular value decomposition (SVD) and iterative Procrustes to use the

composition of two resulting non-identity matrices as the projection matrix of the source language word vector into the target language vector space [10].
- **Bilingual Word Embedding Mappings** (**VecMap**) also uses SVD, but its goal differs from the MUSE and Linear transformation ones. The non-identity matrices resulting from SVD are applied separately to the word vectors of the source and target languages to project them into a new common vector space [2].

In the supervised versions of the MUSE and the VecMap algorithms, matrices for SVD are compiled using a bilingual dictionary.

Several papers dive deeper into the matter of document vectorisation [11,27]. But it is not the focus of this paper. In our experiment, each document was represented as an averaged vector of embeddings of the words included in the text. Some other papers focus on cross-lingual retrieval based on word embedding models [13,24], though they are not dedicated to domain-specific task of academic papers recommendation. Our experiment aims to evaluate different cross-lingual models on academic papers in particular.

3 Experimental Setup

As mentioned in Sect. 1, the goal of the experiment was to find out which approach is more beneficial for the cross-lingual recommendation system task: using pre-aligned cross-lingual models or aligning pretrained monolingual embeddings into a shared vector space yourself. To do this, it was decided to build a bilingual semantic recommendation system for the papers written in English and Russian using cross-lingual embeddings obtained with both approaches. To conduct the experiment, the workflow was organised in three stages:

1. **Preprocessing.** All texts were lemmatised with attaching the part-of-speech (POS) tags to words and deleting words with the functional POS tags. For both languages we used pretrained UDPipe [23] models[4]: the Russian model was trained on the SynTagRus corpus, and the English model was trained on the ParTUT corpus, both with the Universal Dependencies ver. 2.4 up.
2. **Generation of a text representation.** The words were converted to numeric vectors using four cross-lingual word embedding models obtained by two approaches (one model was off-the-shelf and the other three were pairs of monolingual embeddings aligned by the authors). In order to proceed to the next stage, each text vector was composed as the normalised average vector of all words in the text. The methods used to obtain the models will be described in more detail in Sect. 3.2.
3. **Search for the closest papers.** After mapping the vectors to the same space, nearest neighbours for the target text vector were selected by multiplying the vector by a matrix of all text vectors. The dot product of a vector and a matrix row can be interpreted as cosine similarity between them. Since each row of the matrix corresponds to a particular text in the corpus, the closest papers to the target vector were found.

[4] https://ufal.mff.cuni.cz/udpipe/models.

3.1 Data

The RusNLP dataset includes 1,983 papers from three major Russian NLP conferences from their beginning till 2019: Dialogue, AIST, AINL. While the language of the latter two conferences is English, Dialogue also accepts papers in Russian. Their number is decreasing over the years: for the Dialogue, the number of papers in Russian in 2007, 2013, and 2019 is 93, 46, and 25, respectively, which is 97%, 56%, and 40% of the total number of papers in that year; there are no papers in Russian in the AIST proceedings since 2014; AINL has never accepted papers in Russian. Despite this, we can only get a complete picture of the Russian NLP community publications by taking into account papers in all languages. For more information about the corpus, see Table 1 and [4,18].

Table 1. RusNLP corpus statistics

Conference	Since	Texts	Russian	English
Dialogue	2000	1,785	1,424	361
AIST	2012	91	21	70
AINL	2015	96	0	96
Total texts		**1,983**	**1,445**	**527**

3.2 Methods

To perform cross-lingual search for the closest papers, it is necessary to map (to align) monolingual word representations for different languages into a common vector space.

Thus, four methods of making document representations cross-lingual were tested (three of them are described in more detail in the Sect. 2). Among these methods we can naturally distinguish two approaches:

1) Align Russian and English vector spaces, using pretrained monolingual Skip-gram [15] word embedding models[5,6] and a bilingual dictionary with approximately 25,600 word pairs (but only 13,400 pairs are contained in the embedding models, mentioned above) from the Facebook repository[7]. Since embedding models contain lemmatised words with the POS tags attached, the bilingual dictionary was also converted to the same format using UDPipe. So, three alignment methods were applied to the same data. It is important to note that the first method (Translation) does not use any alignment in the strict sense, it is a simple word-by-word translation using only one monolingual embedding model, while the others align two independent vector spaces to create cross-lingual word embeddings.

[5] http://vectors.nlpl.eu/repository/20/3.zip.
[6] http://vectors.nlpl.eu/repository/20/182.zip.
[7] https://github.com/facebookresearch/MUSE#ground-truth-bilingual-dictionaries.

- **Translation** of texts from Russian to English with the dictionary and vectorisation by the English model. This method was chosen as a baseline because it is simple, however is not cross-lingual in its strict sense. Preprocessed Russian texts are translated word-by-word using the bilingual dictionary preprocessed by the same UDPipe model. They are then vectorised by the English word embedding model, thus being mapped to a shared vector space with the original English texts. This method ignores words that are not present in the bilingual dictionary, while other methods can handle them if they are present in embedding models.
- **Linear transformation** of vectors from the Russian model vector space to the English one trained on the bilingual dictionary (hereafter **Linear projection** or **Projection**).
- **VecMap** alignment algorithm, applied to the same Russian and English Skip-gram word embedding models using the same bilingual dictionary.

2) Using pretrained, already pre-aligned in common vector space and therefore cross-lingual word embeddings.
   - **MUSE** aligned word embedding models[8], provided off-the-shelf by Facebook.

In addition to the difference in approaches for obtaining, the models differ in the type of embeddings they are based: the monolingual Skip-grams used for alignment contain lemmatised words with attached POS tags, while the MUSE models are based on fasttext [5] embeddings with unpreprocessed words. All of them are pretrained on the Wikipedia. This study uses only pretrained monolingual vector spaces: this decision was made because our dataset is not big enough to train specialised word embeddings from scratch, although this may have impact on the recommendations quality.

Figure 1 shows the algorithms of all methods of making document representations cross-lingual based on word embeddings.

3.3 Evaluation Setup

To evaluate the quality of search for the most similar papers, annotators with expertise in the field and knowledge of both languages were found through crowdsourcing, provided with guidelines and asked to evaluate the outputs from each method by specifying how many recommended papers are relevant to the target one. Each output sample for a particular target paper was evaluated by three annotators, allowing us to calculate for each method not only the average ratio of relevant recommendations, but also the inter-rater agreement: the Krippendorff's alpha coefficient [8]. 15 annotators took part in the evaluation, but most of it was carried out by the authors of this paper, so the annotation can be trusted. We randomly selected 20 papers in Russian and 20 papers in English from the RusNLP corpus as the target papers, and find the closest five papers for each of

[8] https://github.com/facebookresearch/MUSE#multilingual-word-embeddings.

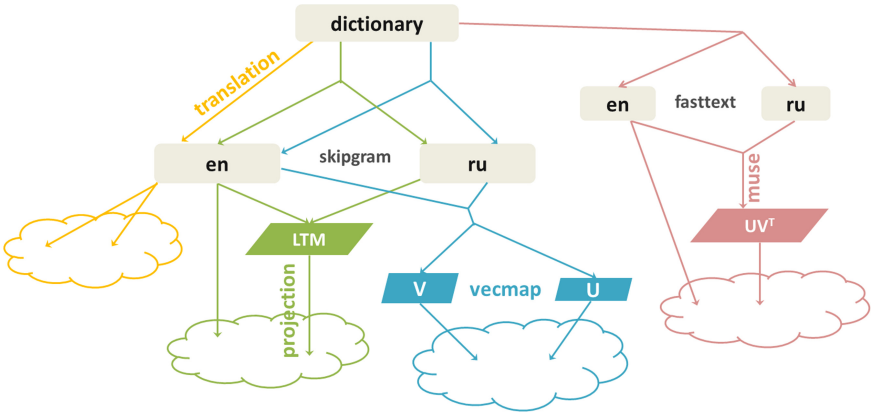


Fig. 1. Methods of making document representations cross-lingual based on word embeddings. *en* and *ru* are pretrained, but not aligned in the common vector space word embedding models; LTM is the abbreviation for the linear transformation matrix; U and V are resulting matrices of singular value decomposition of $Y \cdot X^T$, where Y is the matrix of the English vectors and X is the matrix of the Russian vectors.

them using cross-lingual models, generated with four methods to map vectors of Russian and English words to the same vector space.

To evaluate all resulting models without taking into account the specificity of the RusNLP dataset, they were additionally tested on another dataset. For this purpose, we selected 54 pairs of articles from the Russian and English Wikipedia with parallel titles (that is, they are marked in Wikipedia as written about the same subject in different languages) from several fields: biology, chemistry and medicine; history and culture; flora and fauna; other. For each article it was automatically evaluated whether the article with a parallel title was included into the top-1, top-5, and top-10 recommendations.

4 Results and Discussion

4.1 Quantitative Results

We evaluated the quality of each method using average precision (the average number of relevant recommended papers given by each method). As for recall, it could not be evaluated since the RusNLP dataset is too large for manual verification.

By observing the results presented in Table 2(a), we can tell that compared with Translation (54.5%), Projection (54.5%) and VecMap (54.2%), MUSE has the highest search relevance (58.5%). It is worth pointing out that all methods except MUSE used Skip-gram embeddings with lemmatised and POS tagged words in the model, and MUSE used Fasttext embeddings trained on a non-lemmatized corpus. Though the gap between their results is not that large, only

the values for the first three methods are fully comparable. However, all the results are important from a practical point of view.

Table 2. RusNLP experimental results for target papers in both languages

(a) Average precision

Method	Precision
Translation	54.5
Projection	54.5
VecMap	54.2
MUSE	**58.5**

(b) Inter-rater agreement

Method	Krippendorff's α
Translation	**0.347**
Projection	0.262
VecMap	0.163
MUSE	0.170

Most of the recommendations turned out to be in the language of the target paper; this will be discussed in detail in Subsect. 4.3.

As already mentioned in the Sect. 3.2, the models differ not only in the methods used for obtaining them, but also in the type of word embeddings. It could be noticed, that MUSE, which used fasttex with unpreprocessed words, outperformed other models based on Skip-grams and containing lemmatised words with POS tags.

Within the same alignment method for each target paper we also evaluated the inter-rater agreement of the annotators using the Krippendorff's alpha coefficient [8]. The closer the α is to 1, the higher is the agreement. The values, presented in Table 2(b) are generally low. The highest Krippendorff's α is only 0.347 (Translation), while the α of MUSE, which has the highest average precision, is only 0.17. Such low consistency means that the annotators had very different opinions about the recommendations produced by MUSE. Thus, this method cannot be considered as the best one with full confidence, especially when the absolute difference in average precision with other methods is small. The following difficulties in the annotation process may have caused such a bias in the annotators agreement:

1. **Ambiguity of the guidelines.** The annotating guidelines could have been not clear enough, which may have caused misunderstandings.
2. **Not paper-specific evaluation.** For every five recommendations under the same target, the annotator was asked how many of them were relevant, but not whether each paper is relevant or not.
3. **Size of the annotation forms.** The forms were too long and too time-consuming to fill in, therefore the annotators could fill them out less carefully.

It could be noted that the presence of cross-lingual results did not affect the quality of the evaluation, since the knowledge of both English and Russian was a requirement for the annotators.

4.2 Examples of Relevant Recommendations

In this section, the relevance of examples will be analysed from our point of view, but this analysis is unavoidably subjective, which is proven by the low inter-rater agreement.

Here is a search target example where, according to the annotators, MUSE has outperformed other alignment methods. The title of the target paper is *'Semantic Role Labelling with Neural Networks for Texts in Russian'*. Five recommendations from the approach using MUSE are listed below:

1. *Semantic Role Labelling for Russian Language Based on Russian FrameBank*
2. *Classification Models for RST Discourse Parsing of Texts in Russian*
3. *Exploiting Russian Word Embeddings for Automated Grammeme Prediction*
4. *Methods for Semantic Role Labelling of Russian Texts*
5. *Wear the Right Head: Comparing Strategies for Encoding Sentences for Aspect Extraction*

The first and the fourth recommendations are truly devoted to the semantic role labeling task and obviously can be considered relevant to the target paper. The second and the fifth papers also seem to be appropriate: they are both devoted to the topics closely related to the semantic role labelling task. Moreover, the second paper also evaluates different neural models as the target one. The third paper is dedicated to the applicability of word embeddings to the prediction of classifying grammemes and seems quite irrelevant to the target. Therefore, the estimated precision for this example is 80% which is even higher than the average precision given by the annotators (66.6%).

Another example is the one where the most recommendations were not in the language of the target paper. It has been proposed by the Translation algorithm. The target paper *'Разработка формализма для описания сегментных морфологических процессов в германских языках и его компьютерная реализация'*. The recommendations from the system are the following:

1. *Morphological Analyser and Generator for Russian and Ukrainian Languages*
2. *Part-of-Speech Tagging: the Power of the Linear SVB-Based Filtration Method for Russian Language*
3. *К проблеме лемматизации несловарных словоформ*
4. *Особенности лексико-морфологического анализа при извлечении информационных объектов и связей из текстов естественного языка*
5. *Grammatical Dictionary Generation Using Machine Learning Methods*

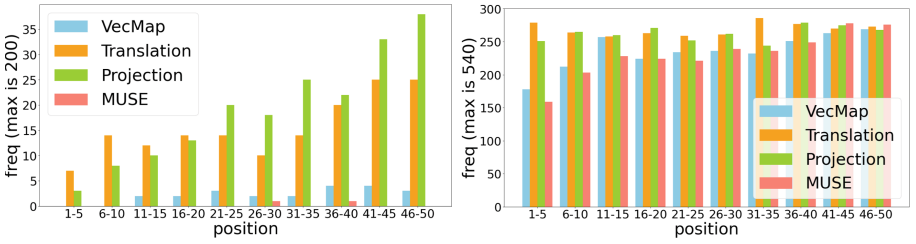
The first and the fourth papers introduce morphological analysers or generators, which is particularly relevant to the target. The other ones can also be appropriate as they are devoted to the tasks closely related to morphology (POS tagging, lemmatisation and automatic grammatical dictionary compilation).

According to the annotators, the average precision for this example is 40%, although we can still detect relevant features in each recommended paper. Evaluation ambiguity may occur as there is a bias in what is considered relevant. The annotation guidelines had a general explanation of relevance but there is still a possibility that it could be interpreted differently.

4.3 Cross-lingual Recommendations

First of all, it should be noted that our main aim was to get the most relevant recommendations regardless of their language. Nevertheless, the difference between recommendations with regards to cross-linguality seem curious and worth some analysis.

As already mentioned, in the vast majority of cases the results were given in the language of the target. This can be partially explained by the specifics of the corpus: papers in Russian (especially from the Dialogue conference) tend to be more often devoted to theoretical aspects of linguistics, while papers about Natural Language Processing and computational linguistics are usually written in English. Thus, for some topics there could be even no papers in the other language to recommend. However, we can say that all results are cross-lingual, but for some methods, recommendations in the other language appear in positions higher than for others. Only Translation and Linear projection managed to yield cross-lingual recommendations among top-5 most similar papers. However, among the recommendations of MUSE and VecMap, there were also cross-lingual ones, but at lower positions. The frequency distribution of cross-lingual recommendations for positions up to 50 is shown in Fig. 2(a). The positions are grouped in bins of five, so the maximum frequency is 200 occurrences: 40 target papers × 5.



(a) RusNLP dataset (b) Wikipedia

Fig. 2. Distribution of cross-lingual recommendations by positions

In other cases, we assume that a large number of words in papers are deleted during vectorisation because they are not present in word embedding models and/or in the bilingual dictionary (more details about the coverage are presented in the Subsect. 4.5). Since the Translation method relies on both resources, more words were lost, and the search for nearest neighbours relied mainly on common words (including words from the examples given in the paper) and frequent terms of syntax and morphology like 'analysis', 'text', 'sentence', 'part', 'form' etc., while specific words are thrown out. The remaining words are translated using the dictionary, vectorised by the same model, and as a result they are converted to the same vector. Since the average representations of the text are generated from mostly identical sets of words, the vectors are very similar. With the other

methods, the texts probably contained many words which were present in only one embedding model (either English or Russian) and had no translations in the other one, so even in the aligned cross-lingual space, the nearest neighbours for these words were words of the same language.

Thus, the reason for the more cross-lingual recommendations in the top positions of the Translation results is that this method has a very limited vocabulary, which makes all papers very close to each other, regardless of the topic. However, we do not yet know the explanation for the difference between the results of Projection and VecMap, which are very similar algorithmically and are trained on the same resources.

It is useful to remember that since our goal is to find the most relevant papers that are not necessarily in a different language, we should not assume that the more cross-lingual results the model offers, the better it is.

4.4 Testing on the Wikipedia Dataset

We conducted an additional experiment to evaluate all resulting models without taking into account the specificity of the RusNLP dataset.

For testing on the Wikipedia, the results are presented in the Table 3. It should be noted that while for the RusNLP dataset the recommendation systems were evaluated manually using precision, the metric for this dataset is recall, since it was automatically evaluated whether the 'correct' article (one with a parallel title) had been included into the top 1, top 5, and top 10 recommendations. The gold standard in this experiment is a list of articles that are marked in Wikipedia as written about the same subject in different languages. The metric for inter-rater agreement (Krippendorff's α) is not relevant for this case, since the evaluation was performed automatically.

Table 3. Wikipedia experimental results for target papers in both languages

Method	Recall@1	Recall@5	Recall@10
Translation	51.85	87.96	95.37
Projection	**56.48**	**91.67**	97.22
VecMap	38.89	85.19	99.07
MUSE	34.26	90.74	**100.00**

In this experiment, MUSE has the lowest recall@1, but the highest recall@10: parallel paper in the other language always occurs in the top-10 recommendations.

On the Wikipedia dataset, all methods turn out to recommend cross-lingual results in high positions, although, as in the experiment on academic papers, Translation and Projection outperform VecMap and MUSE in this respect (Fig. 2(b)). As for MUSE, recommendations in another language are distributed mostly across low positions.

These results are not quite comparable to the actual results on academic papers not only because different metrics (precision and recall) are used, but also since the corpora contain texts of different genres. However, they allow us to get an idea of the quality of all methods without taking into account the specificity of the RusNLP dataset.

So, we can say that the results of the experiment on academic papers were confirmed. MUSE outperformed the other methods as measured by recall@10, which is arguably more important for the recommendation task than recall@1, because it is desirable not to miss any of the relevant papers.

4.5 Analysing Coverage

We also calculated the token and vocabulary coverage for the papers. Table 4 shows percentage of tokens from the text length and the percentage of unique words from the text vocabulary taken into account when vectorising by each model. The values of some models are equal, since the main difference between them is in the vector space aligning methods, not in the model dictionaries. Thus, these results characterise not the methods, but the dictionaries of embedding models used.

Table 4. Coverage (%)

Method	English texts			Russian texts		
	Tokens	Vocab	Dict size	Tokens	Vocab	Dict size
Translation	71.53	63.15	296,630	53.91	47.99	19,118
Projection	71.53	63.15	296,630	**89.30**	**85.57**	248,978
VecMap	71.53	63.15	296,630	**89.30**	**85.57**	248,978
MUSE	**89.30**	**83.21**	200,000	86.58	82.84	200,000

The highest number of words was excluded when vectorising Russian texts with the Translation method, because it used words which had been included in the intersection of the bilingual dictionary and the model dictionary. However, this method had most recommendations in another language in the top-10, and the average precision was the same as for VecMap. At the same time, MUSE, which received the best average precision and gave the least cross-lingual recommendations, had the highest coverage. It can be assumed that the percentage of coverage affects the precision and the amount of cross-lingual recommendations.

5 Conclusions and Future Work

The research question of this paper stemmed from the practical issue of choosing an approach to make document representations cross-lingual for the semantic recommendation system. Hence, the aim was to find the best option for our task between off-the-shelf solution and aligning presumably high-quality pretrained monolingual embedding models on our own. It turned out that the first approach is good enough.

MUSE has the best average precision (58.5%). Other three methods were slightly worse, however, Translation and Projection seem to provide more cross-lingual recommendations than MUSE and VecMap. The fact that most of recommended papers are in the same language as the target one can be explained by algorithmic constraints of all four methods or by the corpus structure.

The results cannot be considered fully consistent as inter-rater agreement is low. It indicates that the annotators were inclined to disagree on the average precision of the four methods. So, further evaluation should be done with several changes: less ambiguous annotating guidelines, paper-specific evaluation of pair relevance, shorter annotation forms. It might be useful to use a ranking task with NDCG as a metric, rather than a binary score, but inter-rated agreement might then be even lower.

Additional testing on a bilingual set of Wikipedia articles demonstrated that models performance may vary depending on the number of recommendations. This experiment confirmed the ambiguity of the results. Nevertheless, in general it showed the same picture as the experiment on the RusNLP dataset.

In practice, as the result of the presented experiments, MUSE outperforms other methods in average precision and demonstrates the largest vocabulary coverage for both languages, therefore can be selected for incorporation into the RusNLP web service until further results are obtained.

Vocabulary coverage is a factor that could affect the results, and we plan to pay attention to it in the future. The results may also depend on the fact that the word embedding models used in the experiment were pretrained on Wikipedia. Presumably, training specialised models on in-domain texts would improve the quality of recommendations. Currently this is not possible to train solely on the RusNLP dataset, since it is too small, but it can be considered as a topic for further experiments. Moreover, while the present experiment was limited to using only word-level embeddings, it would be interesting to apply text-level vectorization methods to the task.

References

1. Artetxe, M., Labaka, G., Agirre, E.: 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), Vancouver, Canada, pp. 451–462. Association for Computational Linguistics, July 2017

2. Artetxe, M., Labaka, G., Agirre, E.: Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, pp. 5012–5019 (2018)
3. Artetxe, M., Ruder, S., Yogatama, D., Labaka, G., Agirre, E.: A call for more rigor in unsupervised cross-lingual learning. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7375–7388. Association for Computational Linguistics, Online, July 2020
4. Bakarov, A., Kutuzov, A., Nikishina, I.: Russian computational linguistics: topical structure in 2007–2017 conference papers. In: Proceedings of Dialogue-2018, Online Papers. ABBYY (2018)
5. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
6. Celli, F., Keizer, J.: Enabling multilingual search through controlled vocabularies: The AGRIS approach. In: MTSR (2016)
7. Klusch, M., Kapahnke, P., Schulte, S., Lécué, F., Bernstein, A.: Semantic web service search: a brief survey. KI - Künstliche Intelligenz **30**, 139–147 (2015)
8. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology. Sage Publications, Thousand Oaks (2018)
9. Kwary, D.A.: A corpus and a concordancer of academic journal articles. Data Brief **16**, 94–100 (2018)
10. Lample, G., Conneau, A., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. In: International Conference on Learning Representations (2018)
11. Lau, J.H., Baldwin, T.: An empirical evaluation of doc2vec with practical insights into document embedding generation. ArXiv abs/1607.05368 (2016)
12. Litschko, R., Glavas, G., Ponzetto, S.P., Vulic, I.: Unsupervised cross-lingual information retrieval using monolingual data only. In: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (2018)
13. Litschko, R., Glavas, G., Vulic, I., Dietz, L.: Evaluating resource-lean cross-lingual embedding models in unsupervised retrieval. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (2019)
14. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation (2013)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
16. Minguillón, J., Lerga, M., Aibar, E., Lladós-Masllorens, J., Meseguer-Artola, A.: Semi-automatic generation of a corpus of Wikipedia articles on science and technology. Profesional De La Informacion **26**, 995–1004 (2017)
17. Moshtaghi, M.: Supervised and nonlinear alignment of two embedding spaces for dictionary induction in low resourced languages. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, pp. 823–832. Association for Computational Linguistics, Novomber 2019
18. Nikishina, I., Bakarov, A., Kutuzov, A.: RusNLP: semantic search engine for Russian NLP conference papers. In: van der Aalst, W.M.P., et al. (eds.) AIST 2018. LNCS, vol. 11179, pp. 111–120. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-11027-7_11

19. Pilehvar, M.T., Camacho-Collados, J.: Embeddings in Natural Language Processing. Morgan and Claypool Publishers (2020)
20. Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual bert? ArXiv abs/1906.01502 (2019)
21. Ruder, S., Vulić, I., Søgaard, A.: A survey of cross-lingual word embedding models. J. Artif. Intell. Res. **65**, 569–631 (2019)
22. Stanković, R., Krstev, C., Obradović, I., Trtovac, A., Utvić, M.: A tool for enhanced search of multilingual digital libraries of e-journals. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, pp. 1710–1717. European Language Resources Association (ELRA), May 2012
23. Straka, M., Straková, J.: CoNLL 2017 shared task: multilingual parsing from raw text to universal dependencies. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada. Association for Computational Linguistics, August 2017
24. Wang, Z., et al.: Estimation of cross-lingual news similarities using text-mining methods. J. Risk Financ. Manage. **11**, 8 (2018)
25. Xu, R., Yang, Y., Otani, N., Wu, Y.: Unsupervised cross-lingual transfer of word embedding spaces. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 2465–2474. Association for Computational Linguistics, October– November 2018
26. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing. IEEE Comput. Intell. Mag. **13**, 55–75 (2018)
27. Zhang, W., Li, Y., Wang, S.: Learning document representation via topic-enhanced LSTM model. Knowl. Based Syst. **174**, 194–204 (2019)
28. Zhang, Y., Gaddy, D., Barzilay, R., Jaakkola, T.: Ten pairs to tag - multilingual POS tagging via coarse mapping between embeddings. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, pp. 1307–1317. Association for Computational Linguistics, June 2016