Clustering of Translation via Topic Modeling

Fedor Krasnov\textsuperscript{1} and Vladimir Lebedev\textsuperscript{2}

\textsuperscript{1} Gazpromneft STC, 75-79 Moika River Emb., Saint Petersburg 190000, Russia, krasnov.fv@gazprom-neft.ru,
\textsuperscript{2} NEICON, Letnikovskaja 4/5, Moscow, Russia, vlebedev@neicon.ru

Abstract. In the present study, the authors investigated structural differences between scientific articles arising from their translation from Russian to English. In the course of the research, a modal topic modelling technique was used. Each document in the assembled collection was presented in two versions: English and Russian. As a result of constructing the topic model, the bimodal matrices $\Phi$ and $\Theta$ were obtained. An analysis of the $\Phi$ matrix showed that the topics could be distinguished according to the degree of correspondence between Russian and English terms when considering words in decreasing order of probability. For 90% of the topics, the English words fully corresponded to the Russian words used. Analysis of the $\Theta$ matrix showed that for 99% of the documents a topic exists having a value greater than 0.95. Thus, the majority of documents are monotopical. Moreover, this majority does not depend on the language of the document. Although the de facto language of scientific articles nowadays is English, a considerable number of scientific works are initially published in scientists' respective native languages and only subsequently translated into English in a more complete and in-depth form. Thus, it is possible to speak about the existence of a bilingual corpus of such documents.

Keywords: Additive regularization, ARTM, Oil&Gas, "Russian English"

1 Introduction

In the field of computer linguistics, there is currently a growing interest in parallel (bilingual) corpora of texts supporting the creation of models for machine translation. For example, the bilingual Proceedings of the Canadian Parliament in English and French form the basis of study [2], while subtitles of films in several languages are analyzed in works [11, 4].

Among the features of such parallel corpora in [12], the language belonging to a particular subject area and the incomplete correspondence of the meanings of translations — i.e. features of translation — are distinguished. The work [8] is devoted to an analysis of features of side-by-side translation in which several levels of parallelism are noted: at the levels of word, phrase, sentence and argument.
The indicated studies [2, 11, 4, 8] focus on identifying pairs of matching sentences, paying great attention to the construction of matches between words. This approach forms an important step in solving the problem of statistical machine translation (SMT), whose formulation more than 50 years ago is described in [14]. Important milestones in the development of translation approaches based on SMT are the creation of Models I and II at IBM’s Thomas J. Watson Research Center during the 1970s. Recently, significant progress has been achieved using the Encoder-Decoder architecture described in [5, 1].

The aim of the present study was to examine the structure of the hidden state, applying Encoder and Decoder not as artificial neural networks, but rather as topical text modelling apparatus using sequential regularization [13].

Different approaches can be taken to the task of translating a scientific article from Russian to English. For example, while some authors use SMT to assist in the process of translation, others opt to rewrite the article in English from scratch. The results of such approaches also differ. In cases where SMT is used, sometimes it is even remarked that such translated articles are written in “Russian English”. Thus, it can be seen that the produced effect is noticeable to human readers. Consequently, an investigation of the criteria used to determine approaches to the translation of articles is not without interest. The research hypothesis examined in this article is as follows:

**Proposition 1.** For a bilingual corpus of scientific articles, it will be possible to automatically identify a pair of scientific articles whose translation was carried out using statistical machine translation.

The study consists of Introduction, Research Methodology, Experiment and Conclusion sections.

2 Research Methodology

In order to provide a formal assessment of the quality of different machine translations, the bilingual evaluation under study (BLEU) metric [10] can be used. However, in our case, this metric is not appropriate, since only one translation variant is studied. A hypothetical approach to testing the proposed research hypothesis is one in which the translation of each article is performed by a native English speaker and then a comparison of the translations is carried out. However, this method would require a significant level of resources. Moreover, this approach would not be capable of evaluating translation options in which the authors had expanded their scientific article during the course of its translation. From the point of view of the BLEU metric, such variants could not be accurately evaluated.

On the other hand, it is possible to translate Russian-language scientific articles into English with the help of the SMT tools such as Moses [6] or Phrasal [3]. The BLEU metric can then be used to measure deviations from the baseline estimation constructed by this means. However, in order to train the SMT, a large bilingual corpus of scientific articles on specific topics would be required.
Given the relatively small dataset, the authors decided to use the method for isolating topics from a text using sequential regularization proposed in [13]. Training of models for the selection of topics was carried out separately on each monolingual corpus of documents. Thereafter, a comparison against the selected topics was performed for each pair of documents.

The essence of topical modelling comprises a decomposition of the vector representation of the text using two matrices ([eq:1]):

\[
p(w|d) = \sum_{t \in T} \varphi_{wt} \theta_{td}
\]

where \( \varphi_{wt} \) is the probability matrix of words \( w \) in each topic \( t \in T \), \( \theta_{td} \) is the probability distribution of topics \( t \) in the document \( d \), and \( p(w|d) \) is the conditional probability of a word \( w \) in the document \( d \). In order to train this model, we used a mechanism for minimizing cross entropy by means of the sequential addition of regularizing terms. Without such regularization, matrices \( \varphi_{wt} \) and \( \theta_{td} \) are of no practical interest.

When there is a variety of \( T \) topics, it is advisable to include topics of two types – main (\( sbj_i \)) and noise (\( nz_j \)). Noise topics may include the introduction and review of scientific sources. For example, as far as the SMT is concerned, although the majority of articles will quote the same fundamental scientific work for the area of knowledge presented in the introduction, the basic structure of articles is likely to vary. For noise topics, the authors carried out regularization with smoothing; conversely, for the main topics, regularization was carried out with sparsification. By this means, the noise level in the main topics was reduced.

The selection of regularization coefficients was carried out according to the method described in study [7]. Following training with matrix regularization, \( \varphi_{wt} \) and \( \theta_{td} \) became sparsified by 80%.

The contents of the matrix \( \varphi_{rus}^{wt} \) and \( \varphi_{eng}^{wt} \) represent the distribution of topics for each document. Thus, in order to analyze the hidden state of the translated text, it is sufficient to analyze the correspondence of topics for different languages. In this case, a problem arises in the translation of the title of one topic from Russian to English. Due to the high sparsity of matrix \( \varphi \), the volume of these works will not be large and can be constructed using an electronic glossary on oil and gas topics.

When mapping \( \varphi_{rus}^{wt} \) and \( \varphi_{eng}^{wt} \), the problem arises of aligning topics for the hidden state. In other words, a single topic in Russian can correspond to one or several topics in English, or remain without correspondence. The complexity of such an alignment in the matrix representation is described by a square matrix having the dimension \((\dim T)^2\).

The obtained hidden matrix representation can be used as a basis for the classification of interconnections between articles written in different languages.
3 Experimental

In order to test the proposed hypothesis, the authors collected a bilingual corpus of 242 articles in English and 242 articles in Russian from the OnePetro.org portal of the international community of oil and gas engineers (SPE). The correspondence of articles in different languages was determined by the DOI index. When creating glossaries, lemmatization was applied and both high- and low-frequency words were discarded. The size of the glossaries for the Russian and English corpora is similar at around words. The training of the model was discontinued when changes in the Perplexity metric

\[ P(D, \Phi, \Theta) = \exp \left( -\frac{1}{n} \sum_{d \in D} \sum_{w \in W} n_{dw} \ln \left( \sum_{t \in T} \phi_{wt} \theta_{td} \right) \right), \quad (2) \]

characterizing the informational entropy of the model started to flatten out.

![Fig. 1: Dependence of the Perplexity metric on the model’s training cycles.](image)

From the figure 1 it can be observed that the entropy of the Russian text is higher than that of the English. A comparison of the Perplexity values for different languages shows agreement with the results published in [9]. The Perplexity metric is strongly dependent on less common words in the glossary.

For regularization, the coefficients \( \mu \) selected in [7] were used. After twelve training iterations with additive regularization, the resulting matrix \( \theta_{dt} \) is represented in the figure 2.

In the figure 2 we can see on the y axis the topics: \( sb_{j0-9} \) – main; and \( n_{z0-1} \) – noise. Due to multidirectional regularization, the space of the main topics is sparsified, while the space of noise topics is smoothed. Examples of
correspondences of topics for Russian and English corpora are shown in the table 1.

From the table 1 it can be seen that topics in English and Russian completely coincide for many $t_i$. This result of visual analysis suggests that the thematic model is tuned to an existing dependency in the data. However, it is necessary to analyze those differences that remain. The density of maximum values $\theta_{td}$ for each document are shown in the figure 3a.

From the figure 3a it can be seen that there are two characteristic classes of documents, which can be divided by the maximum value. Let us consider the documents in more detail, with diametrically opposite values of the maximum: 0.4101 (Document No. 214) and 0.9998 (Document No. 53).

(a) Плотность значений $\theta_{td}$ после обучения модели.

(b) Значения $\theta_{td}$ для разнотипных документов.

Fig. 3: Значения $\theta_{td}$. 

Fig. 2: Matrix of values $\theta_{td}$ following training of the model.
Table 1: Table with topics from matrices $\varphi^{\text{ru}}_{\text{wt}}$ and $\varphi^{\text{en}}_{\text{wt}}$.

| Topic (RU) | Topic (EN) |
|------------|------------|
| реакция, колонна, приток, насос, раствор, обсадный, концентрация | reaction, casing, pump, logging, log, mixture, string |
| долото, управление, дк, наземный, интегрировать, мощность, пара | bit, steam, network, integrated, bcs, run, risk |
| трещина, гри, ячейка, приток, буревой, сетка, раствор | fracture, grid, equation, horizontal well, hydraulic, liner, boundary |
| трещина, напряжение, гри, модуль, геомех-ский, упругий, трещиностойкость | fracture, stress, medium, hydraulic, elastic, geomechanical, fracturing |
| приток, колонна, установка, оборудование, заканчивание, песок, труба | and, esp, completion, inflow, pump, failure, tubing |
| кислотный, воздействие, отложение, кислот, залежь, разрез, карбонатный | acid, treatment, stimulation, carbonate, acid treatment, pilot, seismic |
| гри, трещина, микросейсмический, мигра, проппант, пор; гипст | fracturing, fracture, hydraulic, hydraulic fracturing, proppant, microseismic, port |
| потери, интегрировать, сбор, расход, управление, нагнетательный, пдд | pipeline, integrated, flow rate, condensate, unit, gathering, reservoir pressure |
| раствор, буревой, колонна, буревой раствор, геомех-ский, строит-во, риск | mud, casing, geomechanical, risk, weight, history, stress |
| неопределен, оторочка, рабочий, залежь, адаптация, конденсат, вытеснение | uncertainty, composition, history, condensate, mineral, sample, assessment |
| образец, керн, эксперимент, раствор, частица, фильтрация, заводление | sample, core, cement, experiment, filtration, pore, strength |
| образец, керн, карбонатный, пора, смаываемость, гис, поровый | porous, carbonate, sample, cement, logging, wettability, space |

From the figure 3b, it can be seen that Document No. 214 fits into the single topic $sb_{j6}$, while document No. 53 is distributed across three topics: $sb_{j2}$, $n_{z0}$ and $n_{z1}$. Let us compare the weights of the terms of each of these topics in order to understand how they are correlated.

The figure 4 shows the ten highest values from matrix $\varphi_{\text{wt}}$ for the topics $sb_{j6}$, $sb_{j2}$, $n_{z0}$ and $n_{z1}$. The probabilities of terms for the $sb_{j6}$ topic are much higher than those for $sb_{j2}$, $n_{z0}$ and $n_{z1}$. This observation indicates the different nature of the translation of documents.

Thus, the collection under study contains documents of two different types, which can be distinguished according to the nature of the distribution of topics they comprise. The majority of documents have a single $\theta_{td}$ value greater than 0.95.

4 Conclusion

A bilingual corpus of documents was investigated by the authors. The main objective of the study was to highlight translation features, which, according to the authors' hypothesis, consisted in different approaches to the translation of texts from Russian into English. It was expected that in cases where authors
used SMTs for the translation, there would be a more accurate correspondence between the English and Russian words in the topics than in those cases where the authors performed the translation themselves.

The hypothesis stated by the authors was confirmed by means of an experiment. Translated articles written in "Russian English" can be accurately identified on the basis of the topic model.

The study identified topic clustering in the additive multimodal model for the corpus of bilingual documents. Documents with the "Russian English" type of translation are distinguishable by values \( \varphi_{ut} \) and \( \theta_{td} \) from documents taking a more creative approach to translation.

The technique proposed by the authors in this article confirms the possibility of using regularization strategies of topic models to obtain compact representations of topics that highlight the particular characteristics of a studied collection of documents.

The result obtained by the authors can be used to automate the assessment of the quality of translated scientific texts.

Fig. 4: Values \( \varphi_{ut} \) for two different types of documents.
References

[1] Denny Britz et al. “Massive exploration of neural machine translation architectures”. In: arXiv preprint arXiv:1703.03906 (2017).
[2] Peter F Brown et al. “The mathematics of statistical machine translation: Parameter estimation”. In: Computational linguistics 19.2 (1993), pp. 263–311.
[3] Spence Green, Daniel Cer, and Christopher Manning. “Phrasal: A toolkit for new directions in statistical machine translation”. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. 2014, pp. 114–121.
[4] Einav Itamar and Alon Itai. “Using Movie Subtitles for Creating a Large-Scale Bilingual Corpora.” In: LREC. 2008.
[5] Guillaume Klein et al. “Opennmt: Open-source toolkit for neural machine translation”. In: arXiv preprint arXiv:1701.02810 (2017).
[6] Philipp Koehn et al. “Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding”. In: Final Report of the 2006 JHU Summer Workshop. 2006.
[7] Fedor Krasnov and Oleg Ushmaev. “Exploration of Hidden Research Directions in Oil and Gas Industry via Full Text Analysis of OnePetro Digital Library”. In: International Journal of Open Information Technologies 6.5 (2018), pp. 7–14.
[8] Mehdi Mohammadi and Nasser Ghasem-Aghaei. “Building bilingual parallel corpora based on wikipedia”. In: 2010 Second International Conference on Computer Engineering and Applications. Vol. 2. IEEE, 2010, pp. 264–268.
[9] Will Monroe et al. “Generating bilingual pragmatic color references”. In: arXiv preprint arXiv:1803.03917 (2018).
[10] Kishore Papineni et al. “BLEU: a method for automatic evaluation of machine translation”. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics. 2002, pp. 311–318.
[11] Jörg Tiedemann. “Improved sentence alignment for movie subtitles”. In: Proceedings of RANLP. Vol. 7. 2007.
[12] Jörg Tiedemann. “Parallel Data, Tools and Interfaces in OPUS.” In: Lrec. Vol. 2012, 2012, pp. 2214–2218.
[13] Konstantin Vorontsov and Anna Potapenko. “Additive regularization of topic models”. In: Machine Learning 101.1-3 (2015), pp. 303–323.
[14] Warren Weaver. “Translation”. In: Machine translation of languages 14 (1955), pp. 15–23.