UDC 004.912
DOI: 10.53297/18293336-2025.2-78

THE SENTENCE-LEVEL CROSS-LINGUAL PLAGIARISM DETECTION METHOD FOR ARMENIAN-ENGLISH AND ARMENIAN-RUSSIAN LANGUAGE PAIRS

G.A. Petrosyan[1], R.R. Sahakyan[2]

[1]National Polytechnic University of Armenia
[2]Vanadzor State University named after H. Tumanyan

Recent advances in transformer-based models provide new opportunities for cross-lingual plagiarism detection by projecting sentences from different languages into a shared semantic space. This also applies to low-resourced languages, where they show state-of-the-art results. In this paper, we present a method for sentence-level cross-lingual plagiarism detection for Armenian-English and Armenian-Russian language pairs. We describe both subtasks – source retrieval and sentence-level alignment based on a language-agnostic dual-encoder model. In the first subtask, suspicious Armenian texts are segmented and compared with English and Russian texts with a POS-tagging-based approach to obtain the possible sources. In the second subtask, we apply a transformer-based dual-encoder model for measuring semantic similarity between sentences. We also fine-tune the selected dual-encoder model on both parallel and paraphrased Armenian-English and Armenian-Russian sentence pairs to enhance sensitivity to semantic alignment and paraphrase detection. As a source of paraphrased pairs, we use two different datasets: paraphrased pairs obtained from a parallel corpus and English paraphrased datasets adopted for our language pairs. We applied the proposed method on two publicly available datasets, adopting one of them for our language pairs. On both datasets, the tuned model outperforms the original one in terms of F1. The obtained results show that the proposed method shows good effectiveness compared to existing methods.

Keywords: cross-lingual plagiarism detection, transformer, cross-lingual sentence embeddings, pre-trained models, POS tagging, paraphrase detection.

Introduction. Today, in a multilingual environment, the rich existence of translation tools has further increased this problem and encouraged the growth of cross-lingual plagiarism detection, where content is copied from one language and translated into another [1, p. 354].

The detection of plagiarized text remains complex and challenging, with accurate identification of various types of plagiarism necessitating diverse approaches and techniques [2].

Low-resourced languages create additional difficulties for cross-lingual plagiarism detection. The main challenges include a lack of training data, limited NLP tools and resources, poor embedding alignment, and unique linguistic fea-

tures. The problem is further complicated when plagiarism involves paraphrasing – rewriting the same sentence by changing words or syntax.

Therefore, for low-resourced languages, the cross-lingual plagiarism detection is an especially challenging and urgent task [3].

In this work, we aim to:

- Propose a sentence-level cross-lingual plagiarism detection method for Armenian-English and Armenian-Russian language pairs using a transformer-based dual-encoder model.
- Fine-tune selected dual-encoder model on parallel and paraphrased pairs.
- Evaluate the effectiveness of the proposed method on two publicly available datasets.

Related work. In this section, we present some recent works that use multilingual transformer models for cross-lingual plagiarism detection. In [4], the authors fine-tuned a pre-trained multilingual masked language model XLM-RoBERTa for different language pairs and used it to classify sentence pairs as translation/plagiarism. They achieved state-of-the-art results for French, Russian, and Armenian languages. Earlier, the authors of [5] trained a language-agnostic sentence encoder to detect plagiarized sentence pairs that have few or no lexis in common for the Russian-English language pair. To calculate the similarity of a sentence pair, they combined the ratio of common lexis with the sentence embeddings similarity. The authors of [6] fine-tune a multilingual dual-encoder model LaBSE (Language-agnostic BERT Sentence Embedding) on paraphrased multilingual pairs (PAWS-X) via contrastive learning to detect cross-lingual paraphrases. They use an Additive Margin Softmax Loss by incorporating some ArcFace features and 'mega-batching' with hard negatives, which improves the embedding space for semantically similar sentence pairs. Similarly, the authors of [7] train a bi-directional dual-encoder model with additive margin loss on parallel corpora. In [8], the authors evaluate six multilingual transformer-based models (mBERT, mDistilBERT, XLM-RoBERTa, SBERTMultilingual MiniLM-L12, SBERT Multilingual MPNet, and Distil SBERT Multilingual) for cross-lingual plagiarism detection, on several language pairs including Armenian-English.

Source retrieval. For the source retrieval task, we modified our method presented in [9,10]. The main idea of the method remains the same - comparing texts by nouns, adjectives, and verbs. However, instead of forming one vector for each text, suspicious Armenian texts are split into sentences and form queries, while source texts in English and Russian are split into overlapping fragments. Each query is built on the basis of lemmatized Armenian tokens (nouns, verbs, adjectives, named entities, numbers) translated into English and Russian, respectively. Text fragments in the source languages are also split into lemmatized tokens. We apply field weighting for different types of tokens in queries, and match them with indexed English (Russian) fragments using the Pyserni [11] library and the BM25 model (k1=1.2, b=0.5). For each Armenian sentence, top-n most relevant candidates are selected from English and Russian text indexes. Then the top-n fragments are aggregated into document-level source candidates.

Sentence-level alignment. After the source retrieval stage, we employ the Language-agnostic BERT sentence embedding (LaBSE) model for sentence-level alignment.

Language-agnostic BERT sentence embedding is a pretrained model, which achieves state-of-the-art performance on various bi-text retrieval/mining tasks [12]. It supports 109 languages, including Armenian. In order to improve LaBSE performance for Armenian-English and Armenian-Russian sentence pairs, we fine-tune it on two types of data: parallel translations and paraphrases.

Parallel sentences. As a source of parallel sentences, we used several publicly available corpora obtained from Opus (NLLB, JW300, ParaCrawl-Bonus, MultiCCAligned, QED, TED2020, NeuLab-TedTalks, Wikimedia). Since parallel sentences of the chosen language pairs in observable datasets are not always correct translations, after downloading all the sentences, we performed a multi-stage filtering to select only "high-quality" pairs. We set limits on the minimum and maximum length of sentences (5-50 words and at least 40 characters on each side), on the ratio of lengths between sentences (0.67–1.8), and on the number matches. We also control the proportion of Latin and Cyrillic words (for the corresponding language pairs) in Armenian sentences (no more than 40%), and the absence of URL/Email. After filtering, we get 1.8m pairs for Armenian-English and 1m pairs for Armenian-Russian.

Hard-negatives mining. For hard negatives mining, we use the algorithm presented in [5] with minor changes and an extended filter system. We build embeddings of all English (Russian) sentences via LaBSE and index them with FAISS using the FlatIP index. For each Armenian sentence - h_i, we extract 100 nearest sentences in English (Russian) - $c_j, j = \overline{1,100}$, where its actual translation - t_i, is normally located in one of the top positions. Then the hard negatives are being selected from the candidates that satisfy the following conditions:

$$\begin{cases} \text{sim}(h_i, c_j) \geq \theta_{min}\ , \\ \text{sim}(h_i, c_j) < \text{sim}(h_i, t_i), \end{cases}$$

where function sim (a, b) determines the semantic similarity between embeddings of multilingual sentences a and b computed by LaBSE, and θ_{min} is the minimum allowed threshold of semantic similarity. We additionally compare the selected candidates with the English (Russian) sentence of each pair. In particular, we apply Jaacard by tokens and bigrams, the intersection of character n-grams, and exclude candidates if there are matches of numbers and dates. We also exclude paraphrases by comparing with Normalized Levenshtein Distance for lexical similarity, BertScore for more accurate semantic similarity, stsb-roberta-base as a cross-encoder ceiling, RoBERTa-large NLI model (fine-tuned on Multi-Genre Natural Language Inference) for borderline cases. All the candidates that satisfy the mentioned filters are sorted by semantic similarity computed by LaBSE on the FAISS search level in descending order and 4 of them are selected. Thus, as hard negatives are only selected sentences that are close enough to the Armenian sentence of the selected pair, but are neither an exact nor a paraphrased translation.

We randomly select 5k parallel sentence pairs with 4 hard negatives each from filtered pairs, for each Armenian-English and Armenian-Russian language pairs.

Paraphrases. As a source of paraphrase sentence pairs, we use two different datasets.

1) We use filtered sentence pairs from Opus and generate paraphrased versions of Armenian sentences. At first, each Armenian sentence - h_i, is translated into English - e_i and then paraphrased using Pegasus [1]. All the generated $p_j \in E_i, j = \overline{1,n}$ candidates are compared with e_i according to several criteria, to choose the most optimal one. Then we check semantic similarity between the translated English sentence and the generated candidates by all-MiniLM-L6-v22, and drop candidates for which $\text{sim}(e_i, p_j) < \eta_{sim}$, where η_{sim} is minimum allowed similarity threshold. We also drop candidates that excessively differ in length from the translated sentence, and miss numbers or named entities. To ensure lexical diversity, we calculate bigram Jaccard distance and normalized token-level Levenshtein distance by this combined score:

$$d_{lex}(e_i, p_j) = 0.6\left(1 - J_2(e_i, p_j)\right) + 0.4(TER(e_i, p_j)),$$

where function $J_2(a, b)$ measures how many word bigrams a and b sentences have in common, and $TER(a, b)$ shows the minimum number of token-level edit operations to turn sentence a into sentence b, normalized by sentence length. Then we keep those candidates for which the following condition is true $d_{lex}(e_i, p_j) > \eta_{lex}$, where η_{lex} is the minimum allowed lexical difference. If no candidate meets the above-mentioned conditions, the thresholds are gradually relaxed.

The final paraphrase p_j^* is selected using the following formula:

$$p_j^* = argmax_{p_j \in E_i'}(\alpha d_{lex}(e_i, p_j) + (1 - \alpha)\text{sim}(e_i, p_j)),$$

where E_i' - is the set of all paraphrase candidates of an English e_i sentence, which satisfy all the previous conditions, and α is a weighting factor that controls the ratio of semantic similarity and lexical diversity, and function $\text{argmax}_{a_i \in A} S(a)$ returns the element a_i from the set A, for which function S gets its maximum value. Then the selected paraphrase p_j^* is translated back into Armenian and replaces the original sentence h_i.

We randomly select 5k parallel sentences, paraphrase them, and mine 4 hard negatives according to the above-mentioned methods.

2) We also use Quora Question Pairs (QQP), PAWS-X [13], and STS-B [14]. These datasets contain paraphrased sentence pairs in English. To adopt these datasets into the Armenian-English language pair, we select 5k pairs from all 3 datasets and translate one sentence of each pair into Armenian. To get paraphrased

[1] https://huggingface.co/tuner007/pegasus_paraphrase

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

pairs in Armenian-Russian, we have translated another English sentence of each selected pair into Russian.

For each paraphrased pair from both datasets, we mine 4 hard negatives according to the above-mentioned method of hard-negative mining from the filtered sentences from Opus. To translate sentences, we used the Yandex Cloud Translate API[3].

Fine-Tuning. To fine-tune LaBSE on parallel and paraphrased pairs, we also employ the ArcFace-style Additive Margin Softmax Loss presented in [6]. However, we do not apply mega-batch mining and use both in-batch and formerly mined hard negatives.

For each $\mathbf{h_i}, \mathbf{e_i}$ ($\mathbf{r_i}$) L-2 normalised sentence embeddings for Armenian-English (Armenian-Russian) positive pair (parallel or paraphrased), we apply a fixed angular margin m:

$$cos(\varphi(\mathbf{h_i}, \mathbf{e_i}) + m),$$

where $\varphi(\mathbf{h_i}, \mathbf{e_i})$ is the angle between sentence embeddings $\mathbf{h_i}$ and $\mathbf{e_i}$.

All $\{\mathbf{e_j}\}_{j \neq i}$ in-batch and $\{\mathbf{n}_i^k\}_{k=1}^4$ hard negatives are scaled with s, and hard negatives are also given an additive γ boost.

Following [6], we also employ the following objective function:

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{p}{p + n + \gamma \times q}\right),$$

where $p = e^{s \times cos(\varphi(\mathbf{h_i}, \mathbf{e_i}) + m))}$, $n = \sum_{j \neq i} e^{s \times cos(\varphi(\mathbf{h_i}, \mathbf{e_j}))}$, q = $= \sum_k e^{s \times cos(\varphi(\mathbf{h_i}, \mathbf{n}_i^k))}$.

Additionally, we repeat the same process for English (Russian) sentences against Armenian candidates (with only in-batch negatives) and average the two directions.

We fine-tune LaBSE for 15k positive pairs and three epochs using Additive Margin Scale Loss with parameters: m=0.25, s=20, and γ=1.3.

Evaluation and Results. In order to evaluate the effectiveness of the proposed method, we will apply it to two publicly available datasets, one of which we have adapted to the considered language pairs by translating suspicious texts into Armenian.

First, we will apply our method to the dataset presented in [4]. This dataset contains 400 suspicious documents in 5 different languages, including Armenian, and 120000 source documents in English. After the source retrieval stage for each Armenian suspicious text, we select 100 nearest texts for further detailed analysis. Table 1 presents the F1 scores of the default and tuned LaBSE models on the dataset presented in [4].

[3] https://translate.api.cloud.yandex.net/translate/v2/translate

Table 1

Evaluation results on the dataset presented in [4]

Method	Precision	Recall	F1
Default	0.810	0.752	0.78
Tuned	0.812	0.758	0.784

The second dataset on which we will apply our method is paraplag_v2[4] presented in [5]. It is designed to evaluate the performance of methods for identifying monolingual (Russian) and cross-lingual (Russian-English) text borrowings. It contains two essays. Essays1 is a dataset that consists of copy-pasted/moderately disguised essays, whereas Essays2 contains only heavily disguised essays [5]. In order to apply our method to this dataset, we translated the suspicious texts from both essays into Armenian using the Yandex Cloud Translate API. We release the translated datasets for public access[5]. All the source texts of both Essays1 and Essays2 were put into a 100k reference collection, randomly chosen from the 2019 Wikipedia dump (in English and Russian, respectively). Similar to the previous dataset, after the source retrieval stage, we select 100 nearest source texts for each Armenian suspicious text (in English and Russian, respectively). Tables 2 and 3 present the F1 scores of the default and tuned LaBSE models on both essays presented in [5].

Table 2

Evaluation results on the dataset presented in [5] for the hy-en language pair

Dataset (hy-en)	Method	Precision	Recall	F1
Essays1	**Default**	0.685	0.777	0.728
	Tuned	0.692	0.78	0.733
Essays2	**Default**	0.556	0.512	0.533
	Tuned	0.572	0.514	0.541

Table 3

Evaluation results on the dataset presented in [5] for the hy-ru language pair

Dataset (hy-ru)	Method	Precision	Recall	F1
Essays1	**Default**	0.545	0.831	0.658
	Tuned	0.602	0.804	0.688
Essays2	**Default**	0.463	0.573	0.512
	Tuned	0.494	0.528	0.511

As follows from Tables 1-3, for Armenian-English language pair, the fine-tuning improved both recall and precision and for Armenian-Russian language pair, it improved precision. For both datasets, the optimal similarity threshold was selected after testing the method on dev, in order to maximize the F1 score.

[4] https://plagevalrus.github.io/content/corpora/paraplag_v2.html

[5] https://drive.google.com/drive/folders/1GzkHr72ajmi9dJ9hOz-EFKI--isROkwy

Comparison with other methods. We compare our method with the method presented in [4] on the same dataset (Table 4). They fine-tuned a pre-trained multilingual masked language model XLM-RoBERTa and used for plagiarism detection.

Table 4

Comparison with other method

Method	Precision	Recall	F1
[4]	0.73	0.72	0.73
tuned	**0.812**	**0.758**	**0.784**

As shown in Table 4, the proposed method shows good effectiveness in terms of F1.

In addition to the proposed approach of comparing multilingual sentence embeddings, in future work, we plan to evaluate the similarity between sentences as a weighted sum of semantic, lexical, and structural similarities. The structural component will be based on a previously developed method that can represent a sentence as a Markov chain of word transitions [15].

Conclusion

The article presents a two-level method for sentence-level plagiarism detection for Armenian-English and Armenian-Russian pairs. For sentence-level alignment, a transformer-based dual-encoder model was applied, which was additionally tuned on parallel and paraphrased sentence pairs for both Armenian-English and Armenian-Russian language pairs. The paper also describes the processes of hard negatives mining and generating paraphrases for Armenian sentences. The method was applied to two publicly available datasets. The obtained results show that the tuned model outperforms the original one in terms of F1.

References

1. A Systematic Review of Multilingual Plagiarism Detection: Approaches and Research Challenges / **C. Bouaine, F. Benabbou, Z. Ellaky, et al** // International Journal of Advanced Computer Science and Applications (IJACSA). – 2025. – Vol. 16, no. 8. – P. 354-372. *https://doi.org/10.14569/IJACSA.2025.0160836*
2. **Alshehri M., Beloff N., White M.** AraXLM: New XLM-RoBERTa Based Method for Plagiarism Detection in Arabic Text // Intelligent Computing (SAI 2024), Lecture Notes in Networks and Systems. – 2024. – Vol. 1017. – P. 81-96. *https://doi.org/10.1007/978-3-031-62277-9_6*
3. **Սահակյան Ռ.Ռ., Պետրոսյան Գ.Ա.** Հետազոտական աշխատանքների ինքնատիպության աստիճանի գնահատման համակարգի նախագծում // Հայաստանի ճարտարագիտական ակադեմիայի Լրաբեր. - 2022. - Հատոր 19, N1. - էջ 98-103:
4. **Avetisyan K., Malajyan A., Ghukasyan T., Avetisyan A.** A Simple and Effective Method of Cross-Lingual Plagiarism Detection // arXiv preprint arXiv:2304.01352. - 2023. *https://doi.org/10.48550/arxiv.2304.01352*

5. **Zubarev D., Tikhomirov I., Sochenkov I.** Cross-Lingual Plagiarism Detection Method // Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2021), Communications in Computer and Information Science. – 2022. - Vol. 1620. - P. 207-222. *https://doi.org/10.1007/978-3-031-12285-9_13*
6. **Fedorova I., Musatow A.** Cross-lingual paraphrase identification // arXiv preprint arXiv:2406.15066. – 2024. *https://doi.org/10.48550/arXiv.2406.15066*
7. Improving Multilingual Sentence Embedding using Bi-directional Dual Encoder with Additive Margin Softmax / **Y. Yang, G. Hernandez Abrego, S. Yuan, M. Guo, et al** // Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19). – 2019. – P. 5370-5378. *https://doi.org/10.24963/ijcai.2019/746*
8. **Ter-Hovhannisyan T., Avetisyan K.** Transformer-Based Multilingual Language Models in Cross-Lingual Plagiarism Detection // 2022 Ivannikov Memorial Workshop (IVMEM). – 2022. – P. 72-80. *https://doi.org/10.1109/IVMEM57067.2022.9983968*
9. **Петросян Г.А., Саакян Р.Р., Саакян В.Р.** Разработка метода поиска кандидатов межъязыковых текстовых заимствований на основе разметки частей речи // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. – 2025. – Том 21, No 4. – С. 517–531. *https://doi.org/10.21638/spbu10.2025.405*
10. **Петросян Г.А., Саакян Р.Р.** Применение метода поиска кандидатов межъязыковых заимствований на типологически разных малоресурсных языках // Вестник НПУА: Информационные технологии, электроника, радиотехника. – 2025. – No 1. – С. 70-78. *https://doi.org/10.53297/18293336-2025.1-70*
11. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations / **J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, et al** // Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21). – 2021. – P. 2356-2362. *https://doi.org/10.1145/3404835.3463238*
12. Language-agnostic BERT Sentence Embedding / **F. Feng, Y. Yang, D. Cer, N. Arivazhagan, et al** // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2022. – P. 878–891. *https://doi.org/10.18653/v1/2022.acl-long.62*
13. **Yang Y., Zhang Y., Tar C., Baldridge J.** PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). - 2019. – P. 3687–3692. *https://doi.org/10.18653/v1/D19-1382*
14. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation / **D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, et al** // Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). – 2017. – P. 1-14. *https://doi.org/ 10.18653/v1/S17-2001*
15. **Саакян Р.Р., Шпехт И.А., Петросян Г.А.** Нахождение наличия заимствований в научных работах на основе марковских цепей // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Про-

цессы управления. – 2023. – Том 19, No 1. – С. 43–50. https://doi.org/10.21638/11701/spbu10.2023.104

Received on 12.11.2025.
Accepted for publication on 29.01.2026.

ՀԱՅԵՐԵՆ-ԱՆԳԼԵՐԵՆ ԵՎ ՀԱՅԵՐԵՆ-ՌՈՒՍԵՐԵՆ ԼԵԶՎԱԿԱՆ ԶՈՒՅԳԵՐԻ ԴԵՊՔՈՒՄ ՆԱԽԱԴԱՍՈՒԹՅԱՆ ՄԱԿԱՐԴԱԿՈՒՄ ՄԻՋԼԵԶՎԱԿԱՆ ԳՐԱԳՈՂՈՒԹՅԱՆ ՀԱՅՏՆԱԲԵՐՄԱՆ ՄԵԹՈԴ

Գ.Ա. Պետրոսյան, Ռ.Ռ. Սահակյան

Տրանսֆորմերային մոդելների վերջին նվաճումները միջլեզվական փոխառությունների հայտնաբերման համար ընձեռում են նոր հնարավորություններ՝ տարբեր լեզուներով գրված նախադասությունները ընդհանուր իմաստային տարածության մեջ պրոյեկտելու միջոցով: Սա վերաբերում է նաև սահմանափակ թվային ռեսուրսներով լեզուներին, որտեղ դրանք ցույց են տալիս առաջնակարգ արդյունքներ: Աշխատանքում ներկայացված է նախադասությունների մակարդակում միջլեզվական փոխառությունների հայտնաբերման մեթոդ՝ հայերեն-անգլերեն և հայերեն-ռուսերեն լեզվական զույգերի դեպքում: Աշխատանքում նկարագրված են փոխառությունների որոնման երկու ենթախնդիրները՝ սկզբնաղբյուրների որոնումը և նախադասության մակարդակում փոխառությունների բացահայտումը՝ հիմնված կրկնակի կոդավորիչով մոդելի վրա: Առաջին ենթախնդրում հայերեն տեքստերը բաժանվում են առանձին մասերի և համեմատվում անգլերեն և ռուսերեն տեքստերի հետ՝ խոսքի մասերի նշագրման մեթոդի միջոցով՝ հնարավոր սկզբնաղբյուրները ստանալու համար: Երկրորդ ենթախնդրում կիրառվում է կրկնակի կոդավորիչով տրանսֆորմերային մոդել՝ նախադասությունների միջև իմաստային նմանությունը հաշվարկելու համար: Ընտրված կրկնակի կոդավորիչ մոդելն ուսուցանվում է թարգմանված և վերաձևակերպված հայերեն-անգլերեն և հայերեն-ռուսերեն նախադասությունների զույգերի միջոցով՝ իմաստային համապատասխանեցման և վերաձևակերպումների հայտնաբերման նկատմամբ մոդելի զգայնությունը բարձրացնելու համար: Որպես վերաձևակերպված զույգերի աղբյուր՝ կիրառվում են երկու տարբեր տվյալների բազմություններ՝ զուգահեռ կորպուսներից ստացված վերաձևակերպված զույգեր և ընտրված լեզվական զույգերի համար թարգմանված անգլերեն վերաձևակերպված տվյալների բազմություններ: Առաջարկվող մեթոդը կիրառվել է երկու բաց հասանելիություն ունեցող տվյալների բազմությունների վրա՝ դրանցից մեկը հարմարեցնելով ընտրված լեզվական զույգերին: Երկու տվյալների բազմություններում էլ ուսուցանված մոդելը գերազանցում է սկզբնական մոդելին F1 չափանիշի համատեքստում: Ստացված արդյունքները ցույց են տալիս, որ առաջարկվող մեթոդը ցույց է տալիս բավարար արդյունավետություն՝ համեմատած գոյություն ունեցող մեթոդների հետ:

Առանցքային բառեր. միջլեզվական գրագողության հայտնաբերում, տրանսֆորմեր, նախադասությունների բազմալեզու էմբեդինգներ, նախապես ուսուցանված մոդելներ, խոսքի մասերի նշագրում, վերաձևակերպման հայտնաբերում:

МЕТОД ОБНАРУЖЕНИЯ МЕЖЪЯЗЫКОВОГО ПЛАГИАТА НА УРОВНЕ ПРЕДЛОЖЕНИЙ ДЛЯ АРМЯНО-АНГЛИЙСКОЙ И АРМЯНО-РУССКОЙ ЯЗЫКОВЫХ ПАР

Г.А. Петросян, Р.Р. Саакян

Последние достижения в области моделей на основе трансформеров открывают новые возможности для выявления межъязыковых заимствований путем проецирования предложений из разных языков в общее семантическое пространство. Это также применимо к малоресурсным языкам, где такие модели демонстрируют передовые результаты. В данной статье представлен метод выявления межъязыковых заимствований на уровне предложений для армяно-английской и армяно-русской языковых пар. Описаны обе подзадачи — поиск источников и выравнивание на уровне предложений на основе языково-независимой модели с двойным кодировщиком. В первой подзадаче подозрительные армянские тексты сегментируются и сравниваются с английскими и русскими текстами с использованием подхода, основанного на разметке частей речи, для извлечения возможных источников. Во второй подзадаче применяется модель с двойным кодировщиком на основе трансформеров для измерения семантического сходства между предложениями. Выбранная модель с двойным кодировщиком также дообучается на армяно-английской и армяно-русской параллельных и перефразированных парах, чтобы повысить чувствительность к семантическому выравниванию и выявлению перефразирований. В качестве источника перефразированных пар используются два разных набора данных: перефразированные пары, полученные из параллельного корпуса, и английские наборы данных с перефразированными парами, адаптированные для выбранных языковых пар. Предложенный метод был применен к двум общедоступным наборам данных, адаптировав один из них для выбранных языковых пар. На обоих наборах данных дообученная модель превосходит исходную по показателю F1. Полученные результаты показывают, что предложенный метод демонстрирует достаточную эффективность по сравнению с существующими методами.

Ключевые слова: выявление межъязыкового плагиата, трансформер, межъязыковые эмбеддинги предложений, предобученные модели, разметка частей речи, выявление перефразирования.