УДК 004.912
DOI: 10.53297/18293336-2025.1-70

ПРИМЕНЕНИЕ МЕТОДА ПОИСКА КАНДИДАТОВ МЕЖЪЯЗЫКОВЫХ ЗАИМСТВОВАНИЙ НА ТИПОЛОГИЧЕСКИ РАЗНЫХ МАЛОРЕСУРСНЫХ ЯЗЫКАХ

Г.А. Петросян[1], Р.Р. Саакян[2]

[1]Национальный политехнический университет Армении
[2]Ванадзорский государственный университет им. О. Туманяна

Несмотря на существование множества подходов к обнаружению межъязыковых заимствований и, в частности, к поиску кандидатов, их применение и эффективность для малоресурсных языков остаются слабо изученными. В данной работе представлены результаты применения разработанного авторами метода поиска кандидатов межъязыковых заимствований, который изначально был применён исключительно к армяно-английской языковой паре, на нескольких типологически разных малоресурсных языках, таких как грузинский, греческий, финский и румынский. Основная идея метода состоит в сравнении многоязычных текстов по разным частям речи отдельно, а именно - по существительным, глаголам и прилагательным. Предполагается, что такое сравнение даст лучшие результаты, чем сравнение по всем словам и для других языков, за счет сравнения наиболее информативных лексических единиц. Такое предположение основано на гипотезе, что существительные имеют наибольшую смысловую важность в предложении, за которыми следуют прилагательные и глаголы. Применение метода к языкам, отличающимся друг от друга по морфологической и синтаксической структуре, направлено на подтверждение этой гипотезы и для других языков и демонстрацию обобщаемости предлагаемого метода. Эксперименты в рамках данной работы были проведены над текстами, взятыми из Википедии. В работе также представлены результаты применения метода с использованием распознавания синонимов с помощью WordNet[1] для всех рассматриваемых языковых пар с целью повышения качества результатов. Полученные результаты могут быть полезны для разработки систем обнаружения межъязыковых заимствований, адаптированных под конкретные малоресурсные языки.

Ключевые слова: межъязыковое заимствование, поиск кандидатов, плагиат, обработка естественного языка, уникальность текста, малоресурсные языки.

Введение. При рассмотрении задачи определения степени уникальности текстов, написанных на языках, недостаточно широко представленных в Интернете (далее малоресурсных языках), особую значимость приобретает

[1] Электронный словарь-тезаурус и набор семантических сетей для английского языка.

задача выявления межъязыковых заимствований из-за нехватки потенциальных источников на исходном малоресурсном языке.

Особенно в странах, где научные работы пишутся и публикуются на национальных малоресурсных языках, разработка и внедрение специализированных систем для обнаружения межъязыковых заимствований являются актуальной проблемой [1].

Хотя задача выявления межъязыкового плагиата теоретически может выполняться вручную, объём данных, количество языков и требуемое время делают её практически невыполнимой [2].

Методы автоматического обнаружения плагиата можно разделить на два основных подхода: обнаружение внешнего плагиата и обнаружение внутреннего плагиата [3].

Обнаружение внутреннего плагиата — это задача проверки того, написан ли весь текст, содержащий плагиат (подозрительный текст), одним автором. Обнаружение внешнего плагиата, напротив, заключается в идентификации частей подозрительного текста, заимствованных из оригинальных источников [4].

В отличие от методов обнаружения внешнего плагиата, методы обнаружения внутреннего плагиата не сравнивают подозрительный текст с внешними текстами, они направлены на поиск стилистических изменений, которые могут указывать на возможное заимствование.

Некоторые методы обнаружения плагиата, особенно при обнаружении внутреннего плагиата, ограничивают предварительную обработку до минимума, чтобы не потерять потенциально полезную информацию [5, стр. 67]. Например, методы внутреннего обнаружения обычно не удаляют знаки препинания [6, стр. 13].

Методы поиска кандидатов в многоязычных корпусах текстов применяются в системах для выявления внешнего плагиата.

В результате применения разработанного метода поиска кандидатов на армяно-английской языковой паре можно сделать вывод о справедливости выдвинутой гипотезы: в предложении, написанном на любом языке, наибольшую значимость имеют существительные, за которыми следуют прилагательные и глаголы.

Постановка задачи. Учитывая огромное количество потенциальных исходных текстов, доступных в Интернете, системы выявления межъязыковых заимствований должны отдавать приоритет разработке и применению методов, обеспечивающих быструю и эффективную обработку текстов. Таким образом, этап поиска кандидатов, на котором из большого корпуса тек-

стов выбираются потенциально релевантные тексты, играет важную роль в выявлении межъязыковых заимствований и определении уникальности текста, поскольку он существенно упрощает и ускоряет этап детального анализа исходного текста.

Чтобы показать, что вышесказанное утверждение о частях речи справедливо не только для армянского, но и для других языков, проведем второй эксперимент. Для второго эксперимента увеличим диапазон рассматриваемых языков.

Целями данной работы являются:

- апробация разработанного метода поиска кандидатов для различных языковых пар, чтобы показать, что выдвинутая гипотеза и полученные результаты применимы и к другим языкам;
- повышение эффективности метода за счет распознавания синонимов с помощью Wordnet.

Общее описание метода поиска кандидатов. В данной работе сравнение между подозрительным текстом на малоресурсном языке и оригинальными текстами на английском языке выполняется с помощью отдельных частей речи – существительных, прилагательных и глаголов. На рисунке представлена общая структура работы метода поиска кандидатов.

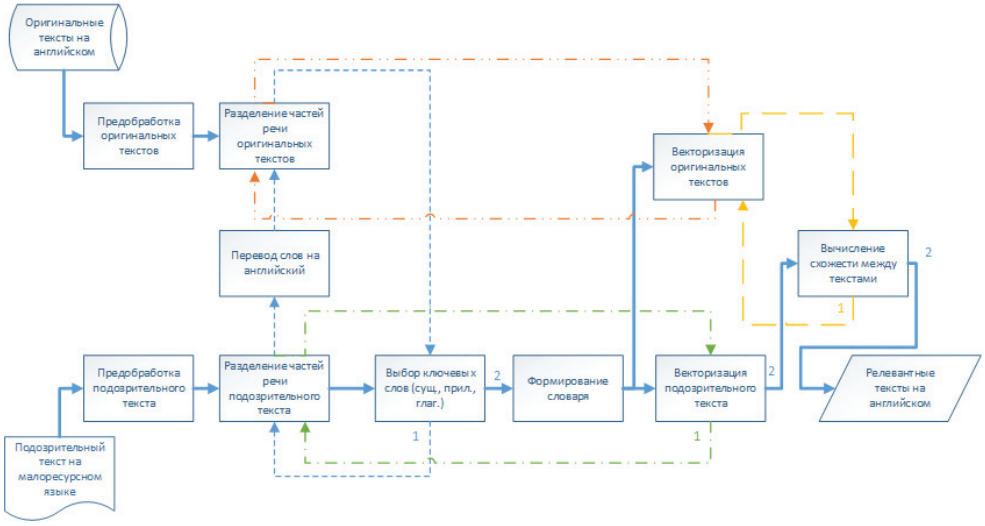

Рис. Общая структура работы метода поиска кандидатов межъязыковых заимствований

В представленной диаграмме блоки, из которых выходят две стрелки, обозначены цифрами, которые указывают порядок их выполнения. Результа-

том работы данного метода являются релевантные тексты на английском языке, из которых в подозрительном тексте на малоресурсном языке предположительно содержатся заимствованные части.

Адаптация метода поиска кандидатов для разных типологических языковых пар. Основные этапы метода остаются неизменными: тексты проходят предварительную обработку, проходя этапы удаления “стоп-слов”, лемматизации и определения принадлежности слов к частям речи. Однако для разных языков используются разные модели, адаптированные под особенности конкретного языка.

Для улучшения результатов метода предлагается при векторизации оригинальных текстов на английском

$$\overrightarrow{V_i} = (\varphi(a_1, V_i), \varphi(a_2, V_i), \dots, \varphi(a_n, V_i))$$

вместо точного совпадения слов использовать метод распознавания синонимов с помощью wordnet:

$$\overrightarrow{V_i} = (\sum_{s_j^1 \in syn(a_1)} \varphi(s_j^1, V_i), \sum_{s_j^2 \in syn(a_2)} \varphi(s_j^2, V_i), \dots, \sum_{s_j^n \in syn(a_n)} \varphi(s_j^n, V_i)),$$

где $syn(a_j)$ – набор всех синонимов слова a_j, взятых из Wordnet, а функция $\varphi(a_j, V_i)$ определяет количество вхождений слова a_j в тексте V_i.

После векторизации сходство текстов определяется с помощью косинуса угла.

Особенность разработанного метода заключается в том, что его эффективность напрямую зависит от ряда факторов, таких как правильное определение принадлежности слов к частям речи, приведение слов к их начальной форме и качество перевода слов.

Эксперименты и результаты

Выбор языковых пар. Для достоверности результатов возьмем типологически разные малоресурсные языки, каждый из которых будет образовывать языковую пару с английским. Выберем два языка из разных ветвей индоевропейской языковой семьи и два из других языковых семей.

В качестве языков для подозрительных текстов выберем следующие: грузинский (картвельская языковая семья), греческий (отдельная ветвь индоевропейской языковой семьи), финский (уральская языковая семья) и румынский (романская подгруппа индоевропейской языковой семьи). А в качестве языка для оригинальных текстов выберем английский язык (германская группа индоевропейской языковой семьи).

Подготовка набора данных. Для создания набора данных в качестве подозрительных и оригинальных текстов были взяты тексты из Википедии. В качестве оригинальных текстов для всех языковых пар будем использовать

10000 текстов из английской Википедии, а в качестве подозрительных текстов - 300 текстов из соответствующих Википедий на всех четырёх языках. Соответствующие тексты из Википедии, как правило, не содержат плагиата и не являются дословными переводами друг друга. Они, в основном, написаны разными авторами, и поскольку они обычно посвящены одной и той же теме, их содержимое, как правило, схоже.

Результаты эксперимента. В табл. 1-3 представлены результаты оценки полноты метода поиска кандидатов для всех рассматриваемых языковых пар при различных частях речи и значениях k ближайших текстов. Помимо трех вышеуказанных частей речи для сравнительного анализа, также сравним тексты по всем словам и параллельно по существительным, прилагательным и глаголам.

Таблица 1

Оценка полноты метода поиска кандидатов для различных языковых пар и частей речи при k=10

Яз. пара	Существительное	Прилагательное	Глагол	Все	Параллельное сравнение
Груз. – англ.	**0.37**	0.35	0.013	0.2	0.5
Греч. – англ.	**0.687**	0.443	0.063	0.44	0.727
Фин. – англ.	**0.323**	0.213	0.11	0.147	0.493
Рум. – англ.	**0.75**	0.613	0.253	0.457	0.85

Таблица 2

Оценка полноты метода поиска кандидатов для различных языковых пар и частей речи при k=20

Яз. пара	Существительное	Прилагательное	Глагол	Все	Параллельное сравнение
Груз. – англ.	**0.433**	0.41	0.027	0.257	0.563
Греч. – англ.	**0.717**	0.51	0.087	0.473	0.787
Фин. – англ.	**0.38**	0.263	0.18	0.203	0.547
Рум. – англ.	**0.793**	0.657	0.323	0.54	0.89

Таблица 3

Оценка полноты метода поиска кандидатов для различных языковых пар и частей речи при k=50

Яз. пара	Существительное	Прилагательное	Глагол	Все	Параллельное сравнение
Груз. – англ.	**0.503**	0.487	0.11	0.317	0.663
Греч. – англ.	**0.79**	0.597	0.16	0.56	0.857
Фин. – англ.	**0.453**	0.333	0.28	0.29	0.683
Рум. – англ.	**0.88**	0.713	0.433	0.623	0.92

Согласно полученным данным, результаты существительных для всех языковых пар превосходят результаты прилагательных, глаголов и всех слов без учета части речи, за которыми следуют прилагательные. Однако наилучшие результаты были получены при сравнении текстов отдельно по трём частям речи параллельно.

Далее представлены результаты сравнения текстов при различных частях речи с использованием распознавания синонимов с помощью WordNet при k=10 (табл. 4).

Таблица 4

Оценка полноты метода поиска кандидатов для различных языковых пар и частей речи при k=10 с использованием распознавания синонимов с помощью WordNet

Яз. пара	Существительное	Прилагательное	Глагол	Все	Параллельное сравнение
Груз. – англ.	0.37	0.35	**0.013**	0.2	0.5
Груз. – англ. (WordNet)	**0.39**	**0.387**	0	**0.297**	**0.573**
Греч. – англ.	0.687	0.443	0.063	0.44	0.727
Греч. – англ. (WordNet)	**0.71**	**0.473**	**0.08**	**0.49**	**0.803**
Фин. – англ.	0.323	**0.213**	**0.11**	**0.147**	0.493
Фин. – англ. (Wordnet)	**0.353**	0.21	0.097	0.137	**0.543**
Рум. – англ.	0.75	0.613	**0.253**	0.457	0.85
Рум. – англ. (Wordnet)	**0.777**	**0.64**	0.22	**0.553**	**0.883**

В большинстве случаев распознавание синонимов позволило улучшить результаты сравнения текстов. Это связано с тем, что распознавание синонимов позволяет идентифицировать слова с одинаковым смыслом, даже если они формально не совпадают.

Заключение

Результаты эксперимента показывают, что для всех рассмотренных малоресурсных языков сравнение текстов по отдельным частям речи, в частности по существительным и прилагательным, дает лучшие результаты, чем сравнение текстов по всем словам. Также использование распознавания синонимов при векторизации текстов способствовало небольшому улучшению результатов.

Стоит также отметить, что при сравнении подозрительных текстов, написанных на всех четырёх языках, с английскими источниками результаты языков индоевропейской языковой семьи значительно превосходят результаты двух языков из других языковых семей.

Таким образом, работа успешно расширяет применимость первичного этапа метода обнаружения межъязыковых заимствований, а именно - поиска кандидатов для нескольких языковых пар. Полученные результаты показывают, что выдвинутая гипотеза справедлива для широкого круга языков.

Литература

1. **Սահակյան Ռ.Ռ, Պետրոսյան Գ.Ա.** Հետազոտական աշխատանքների ինքնատիպության աստիճանի գնահատման համակարգի նախագծում // Հայաստանի ճարտարագիտական ակադեմիայի Լրաբեր. - 2022. - Հատոր 19, N1. - էջ 98-103:
2. **Franco-Salvador M., Rosso P., Montes-y-Gómez M.** A systematic study of knowledge graph analysis for cross-language plagiarism detection // Information Processing & Management. - 2016. - 52(4). - P. 550–570.
3. **Rosso P.** On the risk of cross-language plagiarism for less-resourced languages such as Amazigh // Proceedings of the 4 Conférence International sur les Technologies de l'Information et de la Communication et l'Amazighe. Les ressources langagiéres : Construction et Exploitation. – 2012. – P. 53-70.
4. **Sharjeel M.** Mono- and Cross-Lingual Paraphrased Text Reuse and Extrinsic Plagiarism Detection: PhD Dissertation. - School of Computing and Communications, Lancaster University, 2020. - 214 p.

5. **Al-Smadi M., Jaradat Z., Al-Ayyoub M., Jararweh Y.** Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features // Information Processing & Management. - 2017. - 53(3). - P. 640–652. https://doi.org/10.1016/j.ipm.2017.01.002
6. **Foltýnek T., Meuschke N., Gipp B.** Academic Plagiarism Detection: A Systematic Literature Review // ACM Computing Surveys. - 2019. - 52(6), article 112. – P. 1-42. https://doi.org/10.1145/3345317

Поступила в редакцию 21.04.2025.
Принята к опубликованию 10.07.2025.

ՄԻՋԼԵԶՎԱՅԻՆ ՓՈԽԱՌՈՒԹՅՈՒՆՆԵՐԻ ԱՂԲՅՈՒՐՆԵՐԻ ՈՐՈՆՄԱՆ ՄԵԹՈԴԻ ԿԻՐԱՌՈՒՄԸ ՍԱՀՄԱՆԱՓԱԿ ԹՎԱՅԻՆ ՌԵՍՈՒՐՍՆԵՐՈՎ ՏԻՊԱԲԱՆՈՐԵՆ ՏԱՐԲԵՐ ԼԵԶՈՒՆԵՐԻ ՎՐԱ

Գ.Ա. Պետրոսյան, Ռ.Ռ. Սահակյան

Չնայած միջլեզվային փոխառությունների բացահայտման և, մասնավորապես, թեկնածուների որոնման բազմաթիվ մոտեցումների առկայությանը, դրանց կիրառությունն ու արդյունավետությունը սահմանափակ ռեսուրսներով լեզուների դեպքում մնում են քիչ ուսումնասիրված: Սույն աշխատանքը ներկայացնում է հեղինակների կողմից մշակված միջլեզվային փոխառությունների թեկնածուների որոնման մեթոդի կիրառման արդյունքները, որն ի սկզբանե կիրառվել է բացառապես հայերեն-անգլերեն լեզվական զույգի համար, տիպաբանորեն տարբեր սահմանափակ ռեսուրսներով լեզուների նկատմամբ, ինչպիսիք են վրացերենը, հունարենը, ֆիններենը և ռումիներենը: Մեթոդի հիմնական գաղափարն է բազմալեզու տեքստերը համեմատել առանձին խոսքի մասերով՝ գոյականներով, բայերով և ածականներով: Ակնկալվում է, որ նման համեմատությունն ավելի լավ արդյունքներ կտա, քան ըստ բոլոր բառերի համեմատությունը, մյուս լեզուների համար ևս՝ առավել տեղեկատվական բառերի համեմատության շնորհիվ: Այս ենթադրությունը բխում է այն վարկածից, որ նախադասության մեջ գոյականներն ունեն ամենամեծ իմաստային նշանակությունը, որին հաջորդում են ածականներն ու բայերը: Մեթոդի կիրառումը այն լեզուների վրա, որոնք տարբերվում են մորֆոլոգիական և շարահյուսական կառուցվածքով, ուղղված է այլ լեզուների համար ևս այս վարկածի հաստատմանը և առաջարկվող մեթոդի ընդհանրացման ցուցադրությանը: Այս աշխատանքի շրջանակներում փորձերը իրականացվել են Վիքիպեդիայից վերցված տեքստերի վրա: Աշխատանքում ներկայացված են նաև WordNet-ի միջոցով հոմանիշների ճանաչման մեթոդի կիրառման

արդյունքները դիտարկվող բոլոր լեզուների զույգերի համար՝ նպատակ ունենալով բարելավել արդյունքների որակը: Ստացված արդյունքները կարող են օգտակար լինել միջլեզվային փոխառությունների հայտնաբերման համակարգերի մշակման համար, որոնք հարմարեցված են որոշակի կոնկրետ սահմանափակ ռեսուրսների լեզուներին:

Առանցքային բառեր. միջլեզվական փոխառություն, թեկնածուների որոնում, գրագողություն, բնական լեզվի մշակում, տեքստի ինքնատիպություն, սահմանափակ ռեսուրսներով լեզուներ:

APPLICATION OF THE METHOD OF SEARCHING FOR SOURCES OF INTERLANGUAGE BORROWINGS IN TYPOLOGICALLY DIFFERENT LOW-RESOURCE LANGUAGES

G.A. Petrosyan, R.R. Sahakyan

Despite the existence of many approaches to detecting cross-language borrowings and, in particular, for candidate retrieval, their application and effectiveness for low-resource languages remain poorly studied. This paper presents the results of applying the method for searching cross-language borrowing candidates developed by the authors, which was initially applied exclusively to the Armenian-English language pair, to several typologically different low-resource languages, such as Georgian, Greek, Finnish and Romanian. The main idea of the method is to compare multilingual texts by different parts of speech separately, in particular, by nouns, verbs, and adjectives. It is assumed that such a comparison will produce better results than comparing all words, for other languages too, due to the focus on the most informative lexical units. This assumption comes from the hypothesis that nouns have the greatest semantic importance in a sentence, adjectives and verbs follow. The application of the method to languages that differ in morphological and syntactic structure aims to confirm this hypothesis for other languages too and demonstrate the generalizability of the proposed method. Experiments in this work were conducted on texts taken from Wikipedia. The paper also presents the results of applying the method with synonym recognition via WordNet across all considered language pairs, aiming to improve the result quality. These results can support the development of cross-language borrowing detection systems adapted to specific low-resource languages.

Keywords: cross-language borrowing, candidate retrieval, plagiarism, natural language processing, text uniqueness, low-resource languages.