Р. В. Кузнецова, О. Ю. Бахтеев, Ю. В. Чехович, Методы обнаружения переводных заимствований в больших текстовых коллекциях, *Информ. и её примен.*, 2021, том 15, выпуск 1, 30–41

DOI: 10.14357/19922264210105

МЕТОДЫ ОБНАРУЖЕНИЯ ПЕРЕВОДНЫХ ЗАИМСТВОВАНИЙ В БОЛЬШИХ ТЕКСТОВЫХ КОЛЛЕКЦИЯХ*

Р. В. Кузнецова[1], О. Ю. Бахтеев[2], Ю. В. Чехович[3]

Аннотация: Рассматривается задача обнаружения переводных заимствований. Для решения предлагается использовать моноязыковой подход — свести задачу обнаружения заимствований к одному языку, используя машинный перевод. В связи со спецификой рассматриваемой задачи предлагаемый алгоритм обнаружения должен быть устойчив к неоднозначностям перевода. Предлагается декомпозировать задачу на несколько этапов. Сначала отбираются документы-кандидаты, устойчивость к неоднозначности перевода достигается за счет замены слов на метки кластеров, полученных с помощью дистрибутивной модели. Затем происходит сравнение найденных кандидатов и рассматриваемого документа, для этого используется отображение текстовых фрагментов документов в векторное пространство высокой размерности. Вычислительный эксперимент проводится для языковой пары «русский–английский» на двух выборках — синтетическом корпусе и на статьях из журналов, входящих в Российский индекс научного цитирования (РИНЦ).

Ключевые слова: автоматическая обработка текстов; машинный перевод; глубокое обучение; переводные заимствования; обнаружение переводных заимствований; дистрибутивная семантика

DOI: 10.14357/19922264210105

1 Введение

Проблема некорректных текстовых заимствований актуальна для сферы образования и научных исследований [1]. По материалам исследования [2], проведенного в 2013 г., более 1500 диссертаций по историческим наукам, защищенных в России после 2000 г., содержат значительные заимствования из других диссертаций.

Для задачи обнаружения заимствований в рамках одного языка высокую полноту поиска показывают промышленные инструменты [1], работа которых основана на представлении документов в виде набора перекрывающих друг друга пословных n-грамм (шинглов) [3]. Такой подход позволяет эффективно проводить поиск точных текстовых заимствований, но не позволяет обнаруживать заимствования с большой долей перефразированного текста или со вставками текста, переведенного с другого языка.

Существуют несколько подходов, описывающих проблему поиска переводных заимствований для некоторых пар языков [4, 5], например для пары испанский–английский. Настоящая работа посвящена обнаружению переводных заимствований для пары языков русский–английский. Данная пара нечасто встречается в литературе и не является родственной. Выбор пары языков русский–английский обусловлен преобладанием англоязычных публикаций в интернете и лучшим знанием этого языка по сравнению с другими. Аналогично работам [6, 7] в данной статье предлагается описание алгоритма полного цикла поиска заимствований — сначала ведется поиск документов-кандидатов по внешней коллекции, затем происходит их детальное сравнение с проверяемым документом. Предлагается алгоритм, основанный на моноязыковом анализе документов, схожем с проведенным в работах [8, 9] — проверяемый документ переводится на английский язык с использованием системы машинного перевода с дальнейшим сравнением текстовых фрагментов внутри документов.

В ряде работ, посвященных поиску переводных заимствований, используются дополнительные ресурсы, такие как тезаурусы и онтологии. В работах [4, 5] авторы предлагают использовать базы знаний для извлечения информации о близости между текстами. В работе [5] предлагается алгоритм, основанный на комбинации нейронных сетей и графов знаний. Основной недостаток этого подхода — ресурсоемкость: использование мультиязычных онтологий и баз знаний требует больших вычисли-

*Работа выполнена при поддержке РФФИ (проект 18-07-01441) и Фонда содействия развитию малых форм предприятий в научно-технической сфере (проект 44116).

[1]Московский физико-технический институт, rita.kuznetsova@phystech.edu

[2]Компания Антиплагиат; Московский физико-технический институт, bakhteev@ap-team.ru

[3]Вычислительный центр им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, chehovich@ap-team.ru

тельных мощностей для построения семантических графов для каждого текстового фрагмента, а также сравнения полученных семантических графов.

В данной работе предлагается декомпозиция алгоритма обнаружения переводных заимствований для поиска по большим текстовым коллекциям. Общая схема алгоритма включает следующие шаги.

1. *Машинный перевод* — перевод проверяемого документа на английский язык. Для этого используется система статистического машинного перевода [10].
2. *Поиск документов-кандидатов* — для проверяемого документа находятся наиболее релевантные документы-кандидаты, для этого используется модификация алгоритма шинглов.
3. *Сравнение документов* — текст разбивается на фрагменты, строится отображение каждой фразы в векторное пространство. Для каждого вектора проверяемого документа находятся ближайшие векторы из документов-кандидатов, после чего проводится классификация пар данных векторов на схожие и несхожие пары текстовых фрагментов.

Так как в предлагаемом алгоритме используется моноязыковой анализ заимствований, то задача близка к задаче обнаружения перефразированного текста. Ряд подходов [11–14] к решению этой задачи используют векторные представления фраз, полученные с помощью нейронных сетей глубокого обучения. В работе [13] предлагается нейронный мешок слов (*англ*. Neural Bag-of-Words) и глубокие усредняющие сети (*англ*. Deep averaging networks). В данной статье предлагается использовать выходы нейронной сети как векторные представления текстовых фрагментов для дальнейшего приближенного алгоритма поиска ближайшего соседа [15].

В работе исследуются свойства предлагаемого метода обнаружения переводных заимствований. Проводится анализ моделей глубокого обучения, используемых на этапе сравнения документов, а также составной оптимизируемой функции. Проверка качества предложенного метода проводится как на синтетической выборке, так и на статьях из журналов, входящих в РИНЦ. Проводится анализ ошибок. Предложенный метод поиска заимствований сравнивается с базовым алгоритмом поиска заимствований, основанным на использовании машинного перевода и алгоритме шинглов.

2 Постановка задачи

Пусть заданы коллекции документов на английском языке

$$D_e = \{d_e^j\}_{j=1}^N$$

и русском языке

$$D_r = \{d_r^i\}_{i=1}^M.$$

Документы на русском и английском языке представимы в виде конкатенации текстовых фрагментов:

$$d_e^j = \left[s_{e_1}^j \sqcup \cdots \sqcup s_{e_h}^j\right];\ \ d_r^i = \left[s_{r_1}^i \sqcup \cdots \sqcup s_{r_k}^i\right].$$

Пусть задана выборка

$$\mathcal{D} = \left\{(d_e^l, d_r^l), \mathrm{RL}^l\right\}_{l=1}^L,$$

где каждой паре документов $(d_e^l, d_r^l)_{d_e^l \in D_e, d_r^l \in D_r}$ сопоставлен список пар фрагментов

$$\mathrm{RL} = \left[(s_{e_1}^l, s_{r_1}^l), \ldots, (s_{e_{k(l)}}^l, s_{r_{k(l)}}^l)\right].$$

Для каждой пары $(s_{e_{\tilde{k}}}^l, s_{r_{\tilde{k}}}^l)$ известно, что фрагмент $s_{r_{\tilde{k}}}^l$ является переводом фрагмента $s_{e_{\tilde{k}}}^l$.

Модель f задается как последовательное выполнение функций filter и comparison, где

$$\text{filter: } (d_r^i, D_e)_{d_r^i \in D_r} \to D_e^{\mathrm{retrieved}_i} \subset D_e,$$
$$\text{comparison: } (d_r^i, D_e^{\mathrm{retrieved}_i})_{d_r^i \in D_r} \to \mathrm{RL}^i.$$

Здесь RL^i — список пар фрагментов. Функция filter отвечает за сужение числа документов коллекции, сравниваемых с проверяемым документом, и позволяет проводить дальнейшее более детальное сравнение comparison с использованием ресурсоемких вычислительных алгоритмов, основанных на моделях глубокого обучения.

Качество модели f оценивается с помощью функций Precision и Recall:

$$\mathrm{Precision} = \frac{|(\cup_{l=1}^L \mathrm{RL}^l) \cap (\cup_{i=1}^M \mathrm{RL}^i)|}{|\cup_{i=1}^M \mathrm{RL}^i|},$$
$$\mathrm{Recall} = \frac{|(\cup_{l=1}^L \mathrm{RL}^l) \cap (\cup_{i=1}^M \mathrm{RL}^i)|}{|\cup_{l=1}^L \mathrm{RL}^l|}.$$

Требуется найти функцию f, максимизирующую F1, среднее гармоническое показателей Precision и Recall:

$$\hat{f} = \arg\max_{f \in \mathcal{F}} \mathrm{F1}(f, \mathcal{D}),$$

$$\mathrm{F1} = \frac{2\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}},$$

где $\mathcal{F}$ — заданное семейство моделей.

3 Поиск документов-кандидатов

Одним из алгоритмов поиска документов-кандидатов в задачах обнаружения дословных заимствований и поиска *почти-дубликатов* текста служит алгоритм, основанный на построении инвертированного индекса, в котором каждый документ коллекции представляется набором *шинглов* [3], т. е. набором перекрывающихся n-грамм. Проверяемый документ также разбивается на шинглы, после чего проводится поиск документов по инвертированному индексу с наибольшим совпадением шинглов. В данной работе предлагается обобщение алгоритма шинглов, позволяющее улучшить качество поиска кандидатов в случае обнаружения переводных заимствований.

Предлагается функция filter следующего вида:

$$\begin{aligned}&\text{filter}\left(d_r^i, D_e\right) = \\ &= \underset{D_e' \subset D_e, |D_e'|=k}{\arg\max} \sum_{d_e^j \in D_e'} \sum_{h \in \mathcal{H}(d_r^i)} \mathbf{I}\left[h \in \mathcal{H}(d_e^j)\right] \Big/ \Big(|d_e^{j'} \in D_e : \\ &\qquad h \in \mathcal{H}(d_e^{j'})|^\alpha + \text{const}\Big)\,.\end{aligned}$$

Здесь $\mathcal{H}$ — множество n-грамм документа, упорядоченная последовательность n меток кластеров, где процедура формирования кластеров описана ниже; $\alpha \in \mathbb{R}$; k — оптимизируемый гиперпараметр.

Для уменьшения влияния неоднозначности перевода на поиск документов-кандидатов предлагается заменять слова на соответствующие им метки кластеров:

$$\{x_1, \ldots, x_n\} \to \{\text{class}\,(x_1), \ldots, \text{class}\,(x_n)\} = h\,,$$

где $x_1, \ldots, x_n$ — слова. Кластеры предварительно выделены из текстового корпуса и содержат семантически близкие слова. Для уменьшения неоднозначности перевода перед разбиением на n-граммы предлагается удалять из текста стоп-слова и проводить лемматизацию. Для учета возможных перестановок слов, возникающих после перевода текста, слова внутри каждой n-граммы сортируются в лексикографическом порядке.

В данной работе для получения кластеров используется модель векторного представления слов, основанная на дистрибутивной гипотезе. Кластеризация проводится с использованием косинусной функции расстояния

$$\cos\left(\mathbf{c}_1, \mathbf{c}_2\right) = \frac{\langle \mathbf{c}_1, \mathbf{c}_2 \rangle}{||\mathbf{c}_1||_2 ||\mathbf{c}_2||_2}, \tag{1}$$

где $\mathbf{c}_1$ и $\mathbf{c}_2$ — векторы из одного векторного пространства.

Ниже приведены примеры полученных кластеров:

– [beer, beers, brewing, ale, brew, brewery, pint, stout, guinness, ipa, brewed, lager, ales, brews, pints, cask];
– [brilliant, excellent, exceptional, finest, outstanding, super, terrific].

4 Сравнение документов

Для сравнения найденных документов-кандидатов $D_e^{\text{retrieved}_i}$ и проверяемого документа (d_r^i) используется модель векторного представления фразы — тексты разбиваются на фрагменты и сравниваются соответствующие им векторы. Ниже представлены детали алгоритма сравнения, а также анализ предлагаемой оптимизационной задачи.

4.1 Модель векторного представления фразы

Рассмотрим подробнее этап построения отображения фрагмента в вектор. Пусть каждому слову документа на языке коллекции поставлен в соответствие вектор $\mathbf{v} \in \mathbb{R}^u$ размерности u. Для простоты будем полагать, что все фрагменты на языке коллекции имеют ограниченную длину n_{col}. Тогда моделью векторизации фрагмента будем называть отображение

$$\mathbf{h} : \mathbb{W} \times \mathbb{R}^{u \times n_{\text{col}}} \to \mathbb{R}^u\,,$$

где $\mathbb{W}$ — пространство параметров модели. Объекты из множества $\mathbb{R}^{u \times n_{\text{col}}}$ являются последовательной конкатенацией векторов векторных представлений слов для фрагментов выборки:

$$\mathbf{x} \in [\mathbf{v}_1, \ldots, \mathbf{v}_{n_{\text{col}}}]^{\mathrm{T}}, \mathbf{x} \in \mathbb{R}^{u \times n_{\text{col}}}.$$

Для работы с фрагментами длиной меньше n_{col} определим некоторый вектор, обозначающий пустое слово.

Модель оптимизируется в режиме частичного обучения с учителем. В качестве оптимизируемой функции используется составная функция ошибки, представляющая собой сумму ошибки реконструкции и ошибки отступа:

$$\alpha E_{\text{rec}}\left(\mathbf{X}_{\text{rec}}, \mathbf{w}\right) + (1-\alpha) E_{\text{me}}(\mathbf{X}_{\text{me}}, \mathbf{w}) \to \min_{\mathbf{w} \in \mathbb{W}}, \tag{2}$$

где E_{rec} — ошибка реконструкции; E_{me} — ошибка отступа; $\mathbf{X}_{\text{rec}}$ и $\mathbf{X}_{\text{me}}$ — обучающие выборки; $\mathbf{w}$ — параметры модели; α — настраиваемый гиперпараметр. Рассмотрим подробнее каждое слагаемое функции ошибки.

Первое слагаемое функции ошибки соответствует модели автокодировщика. Пусть задана выборка $\mathbf{X}_{\text{rec}} \subset \mathbb{R}^{u \times n_{\text{col}}}$. Модель $\mathbf{h}$ выступает в качестве функции кодирования информации о выборке $\mathbf{X}_{\text{rec}}$. Пусть также задана вспомогательная функция декодирования $\mathbf{g}$, восстанавливающая исходное векторное представление $\mathbf{x}$ по выходам модели $\mathbf{h}$:

$$\mathbf{r}(\mathbf{x}, \mathbf{w}) = \mathbf{g}(\cdot, \mathbf{w}) \circ \mathbf{h}(\mathbf{x}, \mathbf{w}) \approx \mathbf{x}, \mathbf{x} \in \mathbb{R}^{u \times n_{\text{col}}}.$$

Минимизируемая ошибка реконструкции выглядит следующим образом:

$$E_{\text{rec}}(\mathbf{X}_{\text{rec}}, \mathbf{w}) = \frac{1}{|\mathbf{X}_{\text{rec}}|} \sum_{\mathbf{x} \in \mathbf{X}_{\text{rec}}} \| \mathbf{x} - \mathbf{r}(\mathbf{x}, \mathbf{w}) \|_2^2 . \quad (3)$$

Выбор ошибки реконструкции в качестве оптимизируемой функции можно обосновать, используя результаты статьи [16]. Будем пользоваться результатами, доказанными в работе [16], где было показано, что автокодировщики с регуляризацией специального вида позволяют оценить распределение $p(\mathbf{X})$ объектов, принадлежащих генеральной совокупности.

Теорема 1 [16]. *Пусть p — дифференцируемая плотность вероятности и $\forall \mathbf{x}_i \in \mathbb{R}^{u \times n_{\text{col}}}\ p(\mathbf{x}_i) \neq 0$. Пусть $\mathcal{L}_{\sigma^2}$ — функция потерь вида*

$$\mathcal{L}_{\sigma^2} = \int\limits_{\mathbb{R}^{u \times n_{\text{col}}}} p(\mathbf{x}) \left[\| \mathbf{x} - \mathbf{r}(\mathbf{x}, \mathbf{w}) \|_2^2 + \sigma^2 \left\| \frac{\partial \mathbf{r}(\mathbf{x}, \mathbf{w})}{\partial \mathbf{x}} \right\|_F^2 \right] d\mathbf{x},$$

где $\mathbf{r}$ дважды дифференцируема; $0 \leqslant \sigma \in \mathbb{R}$. Пусть $\hat{\mathbf{w}}$ — оптимум функции $\mathbf{r}$ по параметрам моделей кодирования и декодирования, доставляющий минимум $\mathcal{L}_{\sigma^2}$. Тогда

$$\hat{\mathbf{r}}_{\sigma^2}(\mathbf{x}, \mathbf{w}) = \mathbf{x} + \sigma^2 \frac{\partial \log p(\mathbf{x})}{\partial \mathbf{x}} + o\left(\sigma^2\right), \quad \sigma^2 \to 0.$$

Используя результаты теоремы 1, можно сделать следующее утверждение.

Теорема 2. *Плотность вероятности представима в виде*:

$$\frac{\hat{\mathbf{r}}_{\sigma^2}(\mathbf{x}, \mathbf{w}) - \mathbf{x}}{\sigma^2} \approx -\frac{\partial}{\partial \mathbf{x}} E(\mathbf{x}),$$

где $(\mathbf{x}) = (1/Z)\exp(-E(\mathbf{x}))$, Z — нормировочная константа.

Доказательство.

$$\mathbf{r}_{\sigma^2}(\mathbf{x}, \hat{\mathbf{w}}) = \mathbf{x} + \sigma^2 \frac{\partial}{\partial \mathbf{x}} \log p(\mathbf{x}) + o\left(\sigma^2\right);$$

$$\frac{\mathbf{r}_{\sigma^2}(\mathbf{x}, \hat{\mathbf{w}}) - \mathbf{x}}{\sigma^2} = \frac{\partial}{\partial \mathbf{x}} \log p(\mathbf{x}) + o(1);$$

$$\frac{\mathbf{r}_{\sigma^2}(\mathbf{x}, \hat{\mathbf{w}}) - \mathbf{x}}{\sigma^2} \approx \frac{\partial}{\partial \mathbf{x}} \log p(\mathbf{x}).$$

Представляя $\log p(\mathbf{x})$ в форме $-E(\mathbf{x}) - \log Z$, получим искомое выражение.

Таким образом, при устремлении регуляризатора σ к нулю получается *языковая модель*, т. е. распределение вероятностей на множестве $\mathbf{X}$ — множестве текстовых последовательностей.

Второе слагаемое составной функции ошибки — ошибка отступа [12]. Для оптимизации этой функции ошибки используется выборка $\mathbf{X}_{\text{me}} = \{(\mathbf{x}_i, \mathbf{x}_j)\}$, состоящая из пар объектов:

$$\mathbf{X}_{\text{me}} = \left[\mathbf{X}_{\text{me}}^A; \mathbf{X}_{\text{me}}^B\right] \subset \mathbb{R}^{u \times n_{\text{col}}} \times \mathbb{R}^{u \times n_{\text{col}}};$$

$$E_{\text{me}} = \frac{1}{|\mathbf{X}_{\text{me}}|} \left(\sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathbf{X}_{\text{me}}} \max(0, \delta - c_-) + \max(0, \delta - c_+) \right), \quad (4)$$

где

$$c_- = \cos(\mathbf{h}(\mathbf{x}_i, \mathbf{w}), \mathbf{h}(\mathbf{x}_j, \mathbf{w})) - \cos(\mathbf{h}(\mathbf{x}_i, \mathbf{w}), \mathbf{h}(\mathbf{x}_{i'}, \mathbf{w}));$$

$$c_+ = \cos(\mathbf{h}(\mathbf{x}_i, \mathbf{w}), \mathbf{h}(\mathbf{x}_j, \mathbf{w})) - \cos(\mathbf{h}(\mathbf{x}_j, \mathbf{w}), \mathbf{h}(\mathbf{x}_{j'}, \mathbf{w}));$$

δ — отступ; cos — функция расстояния (1),

$$\mathbf{x}_{i'} = \operatorname*{arg\,max}_{\mathbf{x}_{i'} \in \mathbf{X}^B, \mathbf{x}_{i'} \neq \mathbf{x}_j} \cos(\mathbf{x}_i, \mathbf{x}_{i'});$$

$$\mathbf{x}_{j'} = \operatorname*{arg\,max}_{\mathbf{x}_{j'} \in \mathbf{X}^A, \mathbf{x}_{j'} \neq \mathbf{x}_i} \cos(\mathbf{x}_i, \mathbf{x}_{i'}).$$

Следующая теорема объясняет поведение данного слагаемого при проводимой оптимизации параметров $\mathbf{w}$ модели $\mathbf{h}$.

Теорема 3. *Пусть выполнены следующие условия.*

1. *Задан гиперпараметр $\delta \in (0, 2)$.*
2. *Мощность выборки $|\mathbf{X}_{\text{me}}|$ ограничена следующей величиной*:

$$|\mathbf{X}_{\text{me}}|(|\mathbf{X}_{\text{me}}| - 1) \leq \leq \sqrt{\pi} \frac{\Gamma((u-1)/2)}{\Gamma(u/2)} \left(\int\limits_0^{\arccos(1-\delta)} \sin^{u-2} x\, dx \right)^{-1}. \quad (5)$$

3. *Подвыборки $\mathbf{X}_{\text{me}}^A$ и $\mathbf{X}_{\text{me}}^B$ содержат все элементы в единственном числе, ни один элемент не встречается в обеих выборках.*

Тогда существует непрерывное отображение $\hat{\mathbf{h}}$ из множества векторных представлений слов $\mathbb{R}^{u \times n_{\text{col}}}$ в векторное пространство $\mathbb{R}^u$, доставляющее глобальный минимум функции $E_{\text{me}} = 0$.

Доказательство. Построим отображение $\hat{\mathbf{h}}$ явно. Положим для каждой пары $(\mathbf{x}_1, \mathbf{x}_2)$: $\hat{\mathbf{h}}(\mathbf{x}_1) = \hat{\mathbf{h}}(\mathbf{x}_2)$.

Тогда функция E_{me} выглядит следующим образом с точностью до множителя:

$$E_{\text{me}} = \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathbf{X}_{\text{me}}} \max\left(0, \delta - 1 + \cos\left(\hat{\mathbf{h}}(\mathbf{x}_i), \hat{\mathbf{h}}\left(\mathbf{x}_{i'}\right)\right)\right) + \\ + \max\left(0, \delta - 1 + \cos\left(\hat{\mathbf{h}}(\mathbf{x}_j), \hat{\mathbf{h}}\left(\mathbf{x}_{j'}\right)\right)\right).$$

Область значений функции ограничена снизу нулем, который достигается при выполнении условий:

$$1 - \delta \geq \cos\left(\hat{\mathbf{h}}(\mathbf{x}), \hat{\mathbf{h}}\left(\mathbf{x}'\right)\right)$$

для любой пары $\mathbf{x} \in \mathbf{X}^A_{\text{me}}$, $\mathbf{x}' \in \mathbf{X}^B_{\text{me}}$, $(\mathbf{x}, \mathbf{x}') \notin \mathbf{X}_{\text{me}}$, $(\mathbf{x}', \mathbf{x}) \notin \mathbf{X}_{\text{me}}$. Число пар, описанных выше, в множестве $\mathbf{X}_{\text{me}}$ при выполнении третьего условия теоремы равно $|\mathbf{X}_{\text{me}}|(|\mathbf{X}_{\text{me}}| - 1)$. Назначим значение отображения $\hat{\mathbf{h}}$ для каждой такой пары так, чтобы $\cos(\hat{\mathbf{h}}(\mathbf{x}), \hat{\mathbf{h}}(\mathbf{x}')) \leq 1 - \delta$.

Существование такого отображения следует из задачи о нахождении сферического кода максимального размера для сферы в пространстве размерности u и углом $\arccos(1-\delta)$. В работе [17] представлена нижняя оценка для размерности выборки, удовлетворяющей заданным условиям. Оценка соответствует правой части неравенства (5). Так как выборка $\mathbf{X}_{\text{me}}$ конечна, то для построения непрерывной функции, заданной условиями, описанными выше, можно использовать интерполяционные полиномы, что и требовалось доказать.

Заметим, что предложенное в теореме отображение является непрерывным, поэтому для приближения данного отображения можно использовать нейросетевые модели. По теореме Цыбенко отображения из класса нейросетевых моделей будут приближать непрерывные модели сколь угодно хорошо [18].

Таким образом, составная оптимизируемая функция (2) позволяет получить модель, которая, с одной стороны, обладает обобщающими свойствами, за которые отвечает языковая модель (3), с другой стороны, эффективно разделяет схожие и несхожие фразы из обучающей выборки (4). Гиперпараметр α отвечает за вклад каждого из оптимизируемых слагаемых в данную функцию.

4.2 Классификатор

Для каждого вектора фразы $\mathbf{h}(\mathbf{x}^i_{r_a})$ из проверяемого документа d^i_r находится v ближайших векторов по косинусной функции расстояния (1) для фрагментов из документов-кандидатов $D_e^{\text{retrieved}_i}$, используя метод приближенного поиска ближайшего соседа. Основная цель данной процедуры — сократить число пар фрагментов для классификации для снижения ресурсоемкости этапа сравнения документа.

Для векторного представления пары фрагментов $(\mathbf{h}(\mathbf{x}^j_{e_b}), \mathbf{h}(\mathbf{x}^i_{r_a}))$ рассматривается следующее решающее правило:

$$f_{\text{frag}}\left(\left(\mathbf{h}\left(\mathbf{x}^j_{e_b}\right), \mathbf{h}(\mathbf{x}^i_{r_a})\right)\right) = \\ = \begin{cases} 1, & \text{если } \cos\left(\mathbf{h}(\mathbf{x}^j_{e_b}), \mathbf{h}\left(\mathbf{x}^i_{r_a}\right)\right) > t_1 \\ & \quad \text{и } p\left(\mathbf{h}\left(\mathbf{x}^j_{e_b}\right), \mathbf{h}\left(\mathbf{x}^i_{r_a}\right)\right) > t_2; \\ 0 & \text{иначе,} \end{cases} \quad (6)$$

где p — вероятность классификатора; t_1 — порог косинусной функции расстояния (1); t_2 — минимальный порог вероятности классификатора.

В качестве признаков используется конкатенация разницы по модулю и покомпонентное произведение компонент вектора $[|\mathbf{h}(\mathbf{x}^j_{e_b}) - \mathbf{h}(\mathbf{x}^i_{r_a})|, \mathbf{h}(\mathbf{x}^j_{e_b}) \odot \mathbf{h}(\mathbf{x}^i_{r_a})]$. В качестве классификатора выступает модель случайного леса.

5 Вычислительный эксперимент

Для анализа качества предложенного алгоритма был проведен ряд вычислительных экспериментов как на синтетической выборке [19], так и на реальных коллекциях документов. В данном разделе приводятся детали порождения синтетических выборок и эксперименты, проведенные на них.

5.1 Синтетическая коллекция переводных заимствований

Для порождения переводных заимствований были использованы документы из английской и русской версии сайта Wikipedia. В качестве коллекции документов D_e были использованы 100 тыс. статей из английской версии Wikipedia. В качестве коллекции проверяемых документов D_r использовалась случайная подвыборка документов из русской версии Wikipedia. Для порождения заимствований для каждого документа $d^i_r \in D_r$ применялся следующий алгоритм.

1. Выбрать документы-кандидаты $\{d^j_e\}$ из коллекции D_e. Для уменьшения разброса лексики в документах-кандидатах и проверяемом документе выбор документов-кандидатов проводился из подвыборки 500 наиболее релевантных документов для проверяемого документа d^i_r. Для определения релевантности использовалась tf · idf-мера. Число документов-кандидатов выбиралось случайно от 1 до 10.

2. Выбрать предложения из документов-кандидатов $\{d_e^j\}$ случайным образом и перевести их на русский язык.
3. Заменить случайные предложения из проверяемого документа d_r^i на переведенные предложения из документов-кандидатов. Доля замененных предложений из проверяемого документа d_r^i выбиралась случайно от 20% до 80%.

5.2 Оптимизация параметров рассматриваемых моделей

В качестве модели векторного представления слов использовалась библиотека `fastText` [20], оптимизация параметров которой проводилась на английской версии Wikipedia. Размерность векторного пространства для векторного представления слов и фрагментов была установлена как 100. Для оптимизации модели векторного представления текстовых фрагментов использовался алгоритм AdaDelta с параметрами $\varepsilon = 10^{-6}$, $\mu = 0{,}95$ и L2-регуляризация $\lambda_2 = 10^{-6}$. Для итоговой функции потерь (2) были установлены следующие значения гиперпараметров: $\delta = 0{,}3$; $\alpha = 0{,}1$. Пороги классификатора (6) были подобраны на основе процедуры кросс-валидации: $t_1 = 0{,}6$; $t_2 = 0{,}5$. Для построения кластеров была использована агломеративная кластеризация на векторах слов. В качестве меры близости слов рассматривалась косинусная функция расстояния (1) между соответствующими векторными представлениями. Итоговая модель содержала 30 тыс. кластеров для 777 тыс. слов. В качестве моделей кодирования $\mathbf{h}$ и декодирования $\mathbf{g}$ использовалась рекуррентная модель GRU (gated recurrent unit) [21]. В качестве системы машинного перевода использовался Moses [10], модель которого была была обучена на 18,5 млн параллельных предложений из корпусов Opus [22]. В качестве выборки для минимизации ошибки реконструкции E_{rec} (3) использовались 10 млн предложений из английской версии Wikipedia. Второе слагаемое функции потерь (4) использует информацию о похожих предложениях $\mathbf{X}_{\text{me}} = \{(\mathbf{x}_i, \mathbf{x}_j)\}$. В качестве выборки таких предложений использовались пары параллельных предложений из корпуса OpenSubtitles [22].

5.3 Детали вычислительного эксперимента

Было проведено три эксперимента на синтетических данных.

1. Поиск кандидатов. В данном эксперименте анализировалось качество полученной модели кластеров слов. В качестве базового эксперимента для сравнения рассматривался алгоритм, основанный на шинглах без приведения слов к меткам кластеров.
2. Сравнение фрагментов текста. В данном эксперименте рассматривался случай, когда отбор кандидатов был проведен полностью корректно: Recall@10 $= 1{,}0$. В качестве базового алгоритма также выступал алгоритм, основанный на шинглах: проверяемый документ d_r^i переводился на английский язык. После этого полученный текст проходил лемматизацию и разбивался на множество перекрывающихся 4-грамм. Для учета возможных перестановок слов при переводе слова внутри каждой 4-граммы сортировались. Результатом сравнения двух документов выступало множество совпавших отсортированных 4-грамм.
3. Эксперимент, оценивающий качество всего алгоритма (поиск кандидатов и сравнение фрагментов текста). Данный эксперимент позволял оценить качество представленного алгоритма в целом.

Результаты эксперимента по поиску кандидатов представлены в табл. 1. Представленный алгоритм, основанный на построении кластеров, дает лучшее качество, чем базовый алгоритм, основанный на шинглах.

Результаты экспериментов по сравнению фрагментов текста представлены в табл. 2. Представленный алгоритм показывает точность, сравнимую с точностью базового алгоритма, и полноту, значительно превосходящую полноту базового алгоритма. Точность базового алгоритма объясняется тем, что данный алгоритм учитывает схожесть только почти-дубликатов текста.

В третьем эксперименте, учитывавшем качество представленного алгоритма в целом, были получены следующие показатели: Precision $= 0{,}83$; Recall $= 0{,}79$; F1 $= 0{,}80$.

Таблица 1 Результаты эксперимента по поиску кандидатов

Алгоритм	Recall@10
Базовый	0,93
Представленный	0,95

Таблица 2 Результаты экспериментов по поиску схожих фрагментов текста

Алгоритм	Precision	Recall	F1
Базовый	0,99	0,15	0,26
Представленный	0,93	0,80	0,85

6 Результаты экспериментов на реальной коллекции научных документов

Для апробации представленного алгоритма был проведен эксперимент по поиску переводных заимствований на коллекции документов из электронной библиотеки eLibrary.ru. Данная библиотека содержит научные документы, входящие в РИНЦ. Данный ресурс также содержит дополнительные метаданные для каждого документа: заголовок, авторов документа, язык документа и принадлежность к тематике, соответствующей Государственному рубрикатору научно-технической информации (ГРНТИ). Для апробации алгоритма в качестве проверяемых документов D_r были подготовлены 2,5 млн документов на русском языке.

В качестве коллекции документов D_e использовались документы из английской версии Wikipedia, документы на английском языке из eLibrary.ru и статьи ресурса arXiv.org. Суммарное число полученных документов составило 7,6 млн.

В силу большого числа проверяемых документов D_r для дальнейшего анализа рассматривались документы, содержащие значительное число найденных заимствований. Была получена 21 тыс. документов со значительным числом заимствований. Из них были проанализированы 7,6 тыс. документов, выбранных случайно. Основной целью эксперимента было обнаружение переводных заимствований, когда заимствование произошло из англоязычного документа в русскоязычный документ. В то же время при анализе полученных результатов был выявлен ряд других срабатываний представленного алгоритма, которые были в дальнейшем разделены на несколько типов:

- переводные заимствования — документ содержит заимствования, переведенные с английского языка, выданные за оригинальный текст;
- другие заимствования — заимствования из русскоязычных ресурсов или заимствования, направление которых нельзя определить по датам документов;
- двуязычные статьи — работы одного и того же автора на двух языках;
- самоцитирование — цитирование автором его англоязычной работы;
- цитирование законов — использование формулировок нормативных актов;
- ошибочные срабатывания — ложно-положительные срабатывания представленного алгоритма;
- другое — срабатывания, которые сложно отнести к какой-либо категории из-за нехватки метаданных или плохого качества текстов.

Таблица 3 езультаты экспериментов для коллекции документов eLibrary.ru

Тип	Количество
Переводные заимствования	921
Другие заимствования	2548
Двуязычные статьи	788
Самоцитирование	669
Цитирование законов	1567
Ошибочные срабатывания	507
Другое	698
Всего	7689

Результаты экспериментов представлены в табл. 3. Заметим, что были проанализированы только 36% всех срабатываний алгоритма, поэтому можно предварительно оценить число документов с переводными заимствованиями по всей коллекции в 2,5 тыс., что составляет 0,1% всех документов. Заметим, что результаты были получены в автоматическом режиме и требуют дальнейшей экспертной верификации.

Распределение доли заимствований в проанализированных документах представлено на рис. 1. Средняя доля заимствований составляет 20%.

Для анализа научных тематик, в которых переводные заимствования происходят наиболее часто, были проанализированы документы, отнесенные к типу *переводные заимствования*. Около 70% проанализированных документов были классифицированы по 10 научным рубрикам. Наибольшая часть документов оказалась распределена между рубриками «Экономика. Народное хозяйство. Экономические науки» и «Право. Юридические науки». За-

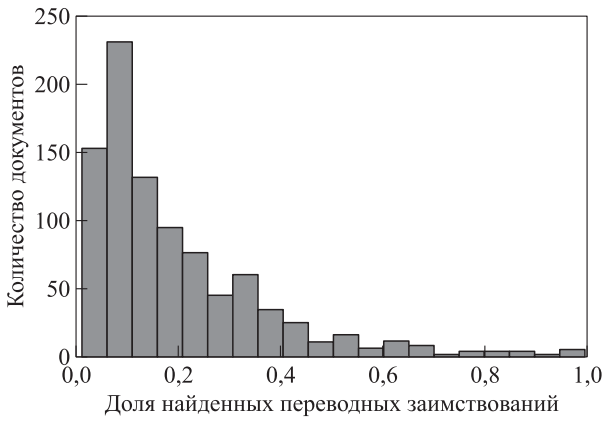

Рис. 1 Гистограмма распределения доли заимствования в тексте

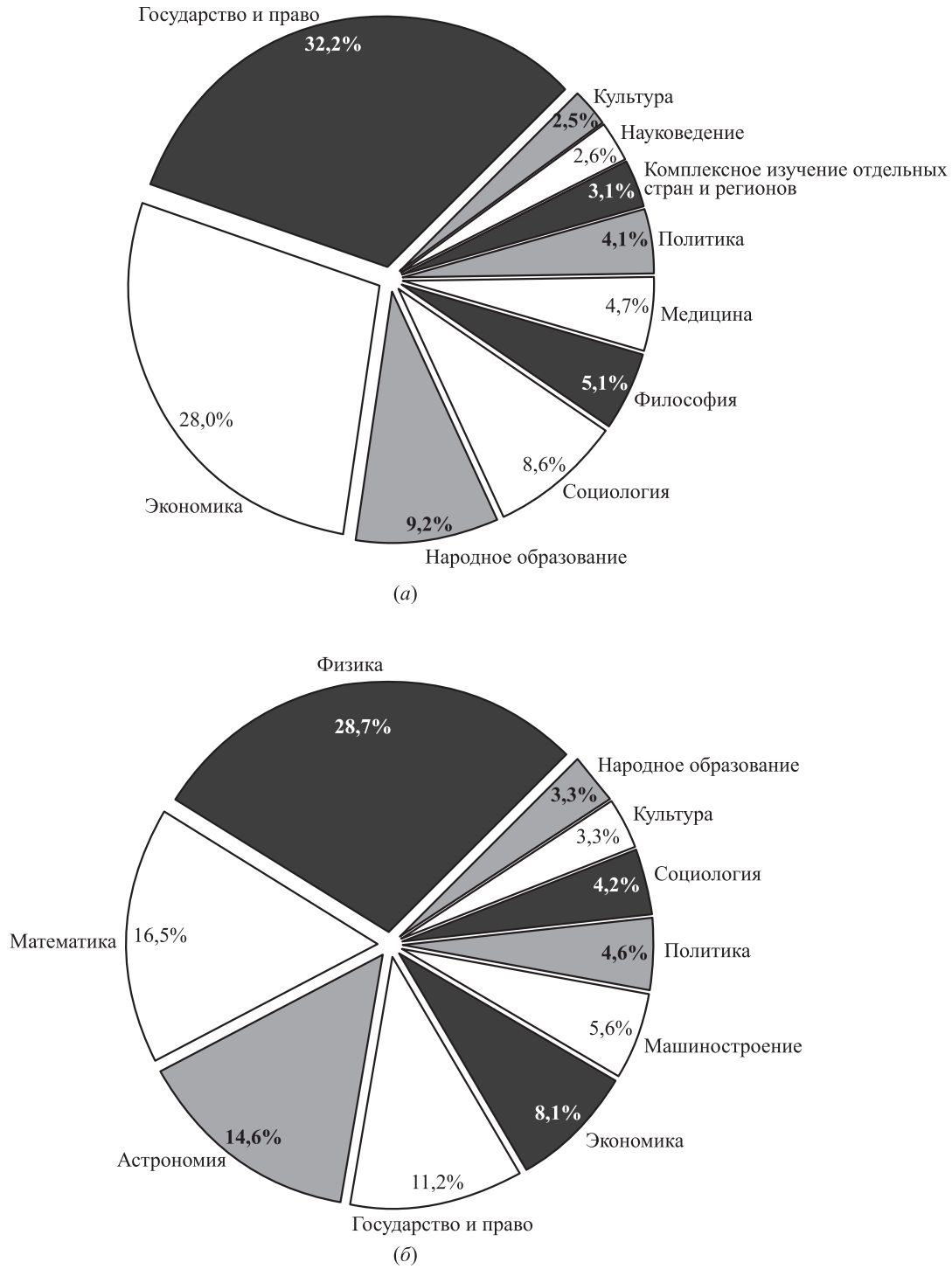

Рис. 2 Распределение заимствований по рубрикам ГРНТИ для типов *переводные заимствования* (*а*) и *двуязычные статьи* (*б*)

метим, что распределение по рубрикам заимствований, отнесенных к типу *двуязычные статьи*, значительно отличается от данного распределения. Диаграммы десяти наиболее представительных рубрик для данных типов срабатываний показаны на рис. 2.

Анализ ложно-отрицательных срабатываний. Для анализа ложно-отрицательных срабатываний представленного алгоритма была проанализирована полнота нахождения двуязычных документов. Оценка полноты была проведена с помощью мета-

данных, полученных из eLibrary.ru. Анализ срабатываний алгоритма показал, что только 85% документов были найдены алгоритмом корректно. Заметим, что представленная оценка полноты является грубой, так как учитывает только полные переводы текстов.

Основная причина ложно-отрицательных срабатываний — низкое качество машинного перевода. Другой проблемой, значительно повлиявшей на качество нахождения двуязычных статей, является используемый алгоритм поиска кандидатов, позволяющий находить только близкие по структуре заимствования. Кроме того, значительная часть проанализированных документов имела некорректную кодировку, что также повлияло на полноту поиска документов.

Анализ ложно-положительных срабатываний. Для анализа ложно-положительных срабатываний были проанализированы вручную 90 документов, отнесенных к типу *ошибочные срабатывания*. Основная проблема ложно-положительных срабатываний состояла в некорректном векторном представлении предложений, содержащих именованные сущности, не встречаемые в обучающей выборке, а также содержащих слова, незнакомые модели машинного перевода. Также было замечено, что алгоритм сравнения документов часто находил общие фразы вида «Работа посвящена следующей проблеме. . .» и т. п. Несмотря на корректность данных срабатываний, общие фразы представленного вида встречаются в большом числе документов и потому не должны рассматриваться как переводные заимствования. Общий процент документов с ложно-положительными срабатываниями составил 7%.

7 Заключение

В работе предложен алгоритм обнаружения переводных заимствований. Предложена декомпозиция алгоритма обнаружения переводных заимствований, позволяющая проводить эффективный поиск заимствований на больших текстовых коллекциях. Проведен анализ предложенного метода обнаружения заимствований, а также составной функции ошибки, используемой для оптимизации модели глубокого обучения. Для анализа качества представленного алгоритма были проведены эксперименты на синтетических данных для пары языков русский–английский. Качество алгоритма было также продемонстрировано на коллекции русскоязычных документов, входящих в РИНЦ. В дальнейшем планируется развитие предложенного алгоритма: использование модели векторного представления предложений для задачи поиска кандидатов и улучшение качества отображения, ставящего в соответствие фразе вектор.

Авторы выражают свою благодарность Г. О. Еременко, ООО «Научная электронная библиотека», за предоставленные материалы.

Литература

1. *Никитов А. В., Орчаков О. А., Чехович Ю. В.* Плагиат в работах студентов и аспирантов: проблема и методы противодействия // Университетское управление: практика и анализ, 2012. Т. 5. С. 61–68.
2. *Khritankov A., Botov P., Surovenko N., Tsarkov S., Viuchnov D., Chekhovich Y.* Discovering text reuse in large collections of documents: A study of theses in history sciences // Artificial Intelligence and Natural Language & Information Extraction, Social Media and Web Search FRUCT Conference. — IEEE, 2015. P. 26–32.
3. *Зеленков И. В., Сегалович И. В.* Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. 9-й Всеросс. научн. конф. RCDL. — Переславль-Залесский: Университет г. Переславля, 2007. С. 166–174.
4. *Franco-Salvador M., Gupta P., Rosso P.* Cross-language plagiarism detection using a multilingual semantic network // European Conference on Information Retrieval / Eds. P. Serdyukov, P. Braslavski, S. O. Kuznetsov, *et al*. — Lecture notes in computer science ser. — Berlin–Heidelberg: Springer, 2013. Vol. 7814. P. 710–713.
5. *Franco-Salvador M., Gupta P., Rosso P., Banchs R.* Cross-language plagiarism detection over continuous-space-and knowledge graph-based representations of language // Knowl.-Based Syst., 2016. Vol. 111. P. 87–99.
6. *Grman J., Ravas R.* Improved implementation for finding text similarities in large collections of data // Notebook papers of CLEF 2011 Labs and Workshops / Eds. V. Petras, P. Forner, P. D. Clough. — Amsterdam, The Netherlands, 2011. Vol. 1177. 6 p. http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-GrmanEt2011.pdf.
7. *Grozea C., Popescu M.* The encoplot similarity measure for automatic detection of plagiarism // Notebook papers of CLEF 2011 Labs and Workshops / Eds. V. Petras, P. Forner, P. D. Clough. — Amsterdam, The Netherlands, 2011. Vol. 1177. http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-GrozeaEt2011.pdf.
8. *Muhr M., Kern R., Zechner M., Granitzer M.* External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system // Notebook papers of CLEF 2010 Labs and Workshops / Eds. M. Braschler, D. Harman, E. Pianta. — Padua, Italy, 2010. Vol. 1176. http://ceur-ws.org/Vol-1176/CLEF2010wn-PAN-MuhrEt2010.pdf.

9. *Bakhteev O., Kuznetsova R., Romanov A., Khritankov A.* A monolingual approach to detection of text reuse in Russian–English collection // Artificial Intelligence and Natural Language & Information Extraction, Social Media and Web Search FRUCT Conference. — IEEE, 2015. P. 3–10.
10. *Koehn P., Hoang Hien, Birch A., et al.* Moses: Open source toolkit for statistical machine translation // 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions Proceedings. — ACL, 2007. P. 177–180.
11. *Tai K., Socher R., Manning C.* Improved semantic representations from tree-structured long short-term memory networks // 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Joint Conference (International) on Natural Language Processing Proceedings. — ACL, 2015. Vol. 1. P. 1556–1566.
12. *Wieting J., Bansal M., Gimpel K., Livescu K.* Towards universal paraphrastic sentence embeddings // arXiv.org, 2015. arXiv:1511.08198 [cs.CL].
13. *Iyyer M., Manjunatha V., Boyd-Graber J. Daume H.* Deep unordered composition rivals syntactic methods for text classification // 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Joint Conference (International) on Natural Language Processing Proceedings. — ACL, 2015. Vol. 1. P. 1681–1691.
14. *Kuznetsova R., Bakhteev O., Ogaltsov A.* Variational learning across domains with triplet information // 3rd Workshop on Bayesian Deep Learning. — Montreal, Canada. http://bayesiandeeplearning.org/2018/papers/65.pdf.
15. *Wang J., Shen H., Song J., Ji J.* Hashing for similarity search: A survey // arXiv.org, 2014. 29 p. arXiv:1408.2927 [cs.DS].
16. *Alain G., Bengio Y.* What regularized auto-encoders learn from the data-generating distribution // J. Mach. Learn. Res., 2014. Vol. 15. No. 1. P. 3563–3593.
17. *Jenssen M., Joos F., Perkins W.* On kissing numbers and spherical codes in high dimensions // Adv. Math., 2018. Vol. 335. P. 307–321.
18. *Cybenko G.* Approximation by superpositions of a sigmoidal function // Math. Control Signal., 1989. Vol. 2. No. 4. P. 303–314.
19. Синтетическая выборка для задачи обнаружения переводных заимствований. https://tiny.cc/cl_ru_en.
20. *Bojanowski P., Grave E., Joulin A., Mikolov T.* Enriching word vectors with subword information // Transactions Association for Computational Linguistics, 2017. Vol. 5. P. 135–146.
21. *Chung J., Gulcehre C., Cho K., Bengio Y.* Empirical evaluation of gated recurrent neural networks on sequence modeling // arXiv.org, 2014. 9 p. arXiv:1412.3555 [cs.NE].
22. *Tiedemann J.* News from OPUS — a collection of multilingual parallel corpora with tools and interfaces // Advances in natural language processing. — Amsterdam/Philadelphia: John Benjamins, 2009. Vol. 5. P. 237–248.

Поступила в редакцию 19.03.2020

METHODS OF CROSS-LINGUAL TEXT REUSE DETECTION IN LARGE TEXTUAL COLLECTIONS

R. V. Kuznetsova[1], O. Yu. Bakhteev[1,2], and Yu. V. Chekhovich[3]

[1]Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141700, Russian Federation
[2]Antiplagiat Co., 42-1 Bolshoy Blvd., Moscow 121205, Russian Federation
[3]A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The paper investigates the cross-lingual text reuse detection problem. The paper proposes a monolingual approach to this problem: to translate the suspicious document into the language of the collection for the further monolingual analysis. One of the major requirements for the proposed method is robustness to the machine translation ambiguity. The further document analysis is divided into two steps. At the first step, the authors retrieve documents-candidates which are likely to be the source of the text reuse. For the robustness, the authors propose to retrieve the documents using word clusters that are constructed using distributional semantics. At the second step, the authors compare the suspicious document with candidates using sentence embeddings that are obtained by deep learning neural networks. The experiment was conducted for the "English–Russian" language pair both on the synthetic data and on the articles included in the Russian Science Citation Index.

Keywords: natural language processing; machine translation; deep learning; cross-lingual text reuse detection; distributional semantics

DOI: 10.14357/19922264210105

Acknowledgments

This research was supported by RFBR (project 18-07-01441) and Foundation for Assitance to Small Innovative Enterprises in Science and Technology (project 44116).

References

1. Nikitov, A. V., O. A. Orchakov, and Y. V. Chekhovich. 2012. Plagiat v rabotakh studentov i aspirantov: problema i metody protivodeystviya [Plagiarism in papers of students and graduate students: The problem and methods of counteraction]. *Universitetskoe upravlenie: praktika i analiz* [University Management: Practice and Analysis] 5:61–68.
2. Khritankov, A. S., P. V. Botov, N. S. Surovenko, S. V. Tsarkov, D. V. Viuchnov, and Y. V. Chekhovich. 2015. Discovering text reuse in large collections of documents: A study of theses in history sciences. *Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference Proceedings*. IEEE. 26–32.
3. Zelenkov, I. V., and I. V. Segalovich. 2007. Sravnitel'nyy analiz metodov opredeleniya nechetkikh dublikatov dlya Web-dokumentov [Comparative analysis of methods for determining fuzzy duplicates for Web-documents]. *Tr. 9-y Vseross. nauchn. konf. "Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii"* [9th All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Electronic Collections" Proceedings]. Pereslavl-Zalessky: Pereslavl-Zalessky University. 166–174.
4. Franco-Salvador, M., P. Gupta, and P. Rosso. 2013. Cross-language plagiarism detection using a multilingual semantic network. *European Conference on Information Retrieval*. Eds. P. Serdyukov, P. Braslavski, S. O. Kuznetsov, *et al*. Lecture notes in computer science ser. Berlin–Heidelberg: Springer. 7814:710–713.
5. Franco-Salvador, M., P. Gupta., P. Rosso, and R. E. Banchs. 2016. Cross-language plagiarism detection over continuous-space-and knowledge graph-based representations of language. *Knowl.-Based Syst.* 111:87–99.
6. Grman, J., and R. Ravas. 2011. Improved implementation for finding text similarities in large collections of data. *Notebook papers of CLEF 2011 Labs and Workshops*. Eds. V. Petras, P. Forner, and P. D. Clough. 1177. 6 p. http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-GrmanEt2011.pdf (accessed January 18, 2021).
7. Grozea, C., and M. Popescu. 2011. The encoplot similarity measure for automatic detection of plagiarism. *Notebook papers of CLEF 2011 Labs and Workshops*. Eds. V. Petras, P. Forner, and P. D. Clough. Amsterdam, The Netherlands. 1177. Available at: http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-GrozeaEt2011.pdf (accessed January 18, 2021).
8. Muhr, M., R. Kern, M. Zechner, and M. Granitzer. 2010. External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system. *Notebook paper of CLEF 2010 Labs and Workshops*. Eds. M. Braschler, D. Harman, and E. Pianta. Padua, Italy. 1176. Available at: http://ceur-ws.org/Vol-1176/CLEF2010wn-PAN-MuhrEt2010.pdf (accessed January 18, 2021).
9. Bakhteev, O., R. Kuznetsova, A. Romanov, and A. Khritankov. 2015. A monolingual approach to detection of text reuse in Russian–English collection. *Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference Proceedings*. IEEE. 3–10.
10. Koehn, P., Hien Hoang, A. Birch, *et al*. 2007. Moses: Open source toolkit for statistical machine translation. *45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions Proceedings*. ACL. 177–180.
11. Tai, K. S., R. Socher, and C. D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *53rd Annual Meeting of the Association for Computational Linguistics and the 7th Joint Conference (International) on Natural Language Processing Proceedings*. ACL. 1:1556–1566.
12. Wieting, J., M. Bansal, K. Gimpel, and K. Livescu. 2015. Towards universal paraphrastic sentence embeddings. 19 p. Available at: https://arxiv.org/abs/1511.08198 (accessed January 18, 2021).
13. Iyyer, M., V. Manjunatha, J. Boyd-Graber, and H. Daumé. 2015. Deep unordered composition rivals syntactic methods for text classification. *53rd Annual Meeting of the Association for Computational Linguistics and the 7th Joint Conference (International) on Natural Language Processing Proceedings*. ACL. 1:1681–1691.
14. Kuznetsova, R., O. Bakhteev, and A. Ogaltsov. 2018. Variational learning across domains with triplet information. *3rd Workshop on Bayesian Deep Learning Proceedings*. Available at: http://bayesiandeeplearning.org/2018/papers/65.pdf (accessed January 18, 2021).
15. Wang, J., H. T. Shen, J. Song, and J. Ji. 2014. Hashing for similarity search: A survey. 29 p. Available at: https://arxiv.org/abs/1408.2927 (accessed January 18, 2021).
16. Alain, G., and Y. Bengio. 2014. What regularized auto-encoders learn from the data-generating distribution. *J. Mach. Learn. Res.* 15(1):3563–3593.
17. Jenssen, M., F. Joos, and W. Perkins. 2018. On kissing numbers and spherical codes in high dimensions. *Adv. Math.* 335:307–321.
18. Cybenko, G. 1989. Approximation by superpositions of a sigmoidal function. *Math. Control Signal.* 2(4):303–314.
19. Sinteticheskaya vyborka dlya zadachi obnaruzheniya perevodnykh zaimstvovaniy [Synthetic dataset for the cross-lingual text reuse detection problem]. Available at: https://tiny.cc/cl_ru_en (accessed January 18, 2021).

20. Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions Association for Computational Linguistics* 5:135–146.
21. Chung, J., C. Gulcehre, K. Cho, and Y. Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. 9 p. Available at: https://arxiv.org/abs/1412.3555 (accessed January 18, 2021).
22. Tiedemann, J. 2009. News from OPUS – a collection of multilingual parallel corpora with tools and interfaces. *Advances in natural language processing*. Amsterdam/Philadelphia: John Benjamins. 5:237–248.

Received March 19, 2020

Contributors

Kuznetsova Rita V. (b. 1990) — PhD student, Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141701, Russian Federation; rita.kuznetsova@phystech.edu

Bakhteev Oleg Yu. (b. 1993) — assistant professor, Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141701, Russian Federation; Head of Research Department, Antiplagiat Co., 42-1 Bolshoy Blvd., Moscow 121205, Russian Federation; bakhteev@ap-team.ru

Chekhovich Yury V. (b. 1976) — Candidate of Science (PhD) in physics and mathematics, Head of Department, A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation; chehovich@ap-team.ru