Методы кросс-языкового поиска тематически похожих нормативно-правовых документов на основе машинного обучения *

В. В. Жебель[I], Д. А. Девяткин[II], Д. В. Зубарев[II], И. В. Соченков[II,III,IV]

[I]Общество с ограниченной ответственностью «Технологии системного анализа», Москва, Россия
[II]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Россия
[III]Университет Иннополис, Казань, Россия
[IV]Институт системного программирования им. В.П. Иванникова Российской академии наук, Москва, Россия

Аннотация. Необходимость изучения мирового опыта для изменения законодательства и нормотворчества вызывает потребность в инструментах информационного поиска нормативно-правовых документов, написанных на разных языках. Одним из аспектов информационного поиска является выявление тематически похожих документов по заданному эталону. В этом контексте возникает важная задача кросс-языкового поиска, когда пользователь информационной системы задает эталонный документ на одном языке, а поисковая выдача содержит релевантные документы на других языках. В настоящем исследовании рассмотрены различные подходы к решению этой задачи: от использования коллекций-медиаторов до более современных методов, опирающихся на дистрибутивную семантику. В качестве тестовой коллекции была использована электронная библиотека ООН, содержащая как оригиналы документов на английском языке, так и их переводы на русский.

Ключевые слова: кросс-языковой поиск документов, дистрибутивная семантика, информационный поиск нормативно-правовых документов.

DOI 10.14357/20718594220203

Введение

Мир стремительно меняется – в результате взрывного технологического прогресса в обиходе людей появляются совершенно новые предметы и явления, юридический статус которых, зачастую, никак не оформлен, что приводит к различным коллизиям. Так, например, одной из задач, сформулированных в указе Президента РФ от 07.05.2018 «О национальных целях и стратегических задачах развития Российской Федерации на период до 2024 года», является создание системы правового регулирования цифровой экономики, основанного на гибком подходе в каждой сфере, а также внедрение гражданского оборота на базе цифровых технологий. Важным этапом при решении подобных задач является изучение мирового опыта и исследование лучших практик. К этому же классу задач относятся вопросы сопоставления тех или иных норм права в разных странах, являющиеся предметом исследования сравнительного правоведения как отдельного раздела юридических наук.

Следует отметить, что существующий объем документов в правовой сфере делает затруднительной их ручную обработку при решении задач юридической компаративистики. Кроме

* Исследование выполнено при финансовой поддержке РФФИ в рамках проекта №18-29-16172.

✉ *Жебель Владимир Викторович. E-mail: zhebel@isa.ru*

того, в разных странах законодательство написано на разных языках с применением специфичной для каждой страны терминологии. Это накладывает дополнительные требования к квалификации эксперта-правоведа, и требует от него систематического, непрерывного изучения нормотворческих инициатив в России и зарубежных странах.

Все это делает задачу кросс-языкового информационного поиска актуальной, а ее решение – востребованным. Поскольку составление подходящего поискового запроса потребует от исследователя глубокого знания особенностей терминологии в языке рассматриваемой страны, одним из лучших решений будет поиск тематически похожих на заданный эталонный документ.

В статье будут рассмотрены различные подходы к решению проблемы кросс-языкового информационного поиска. После этого мы подробнее остановимся на модификациях этих методов для решения именно нашей задачи – поиск нормативно-правовых документов. Далее будет представлена методика и результаты проведенного экспериментального исследования, а затем рассмотрено дальнейшее развитие и практическое применение.

1. Обзор исследований в области кросс-языкового информационного поиска

Суть самой задачи кросс-языкового информационного поиска заключается в том, что пользователь составляет поисковый запрос на одном языке, а результаты ищутся среди документов, написанных на другом. Такой эффект достигается благодаря использованию дополнительной процедуры преобразования, обеспечивающей связь между языками.

Одним из базовых подходов к решению проблемы является перевод поискового запроса на целевой язык с использованием систем машинного перевода, основанных на тезаурусах, онтологиях, базах знаний и т.п. [1, 2] Следует отметить, что из-за неоднозначности перевода (одно и то же слово может быть переведено по-разному – в зависимости от контекста), имеет смысл использовать предметно-ориентированные тезаурусы, в противном случае качество поиска сильно упадет. Такой подход реализован в большом количестве систем: Lingvo, MultiLex, Eckado, RetrievalWare WordNet [3]. Кроме того, можно смотреть на задачу с другой стороны – вместо перевода поискового запроса, можно переводить документы на язык запроса [4]. В [5] сравниваются эти два подхода, и отмечается возможность их комбинирования для повышения эффективности.

Другой интересный подход, описанный в [6], заключается в построении единого семантического пространства с использованием LSI (Latent Semantic Indexing).

Также в последнее время большое распространение получают модели, основанные на нейронных сетях – двуязычный автоэнкодер [7], сиамские сети [8], нейронный машинный перевод [9] и т.п.

В статье мы сконцентрируемся на методах, применимых при поиске тематически схожих документов – в качестве поискового запроса подается целый документ на одном языке, а поисковая выдача состоит из тематически похожих на него документов на другом языке. Рассмотрим подробнее некоторые из подходов.

2. Модификации методов кросс-языкового поиска нормативно-правовых документов

В этом разделе мы рассмотрим методы кросс-языкового информационного поиска, модифицированные для решения задачи автоматизации поиска нормативно-правовых документов.

Для численной оценки качества работы методов, а также для обучения моделей, необходимо выбрать подходящий корпус параллельных документов. Мы решили использовать для этих целей документы из цифровой библиотеки ООН[1] на русском и английском языках. В результате получили выровненный набор из 33 тысяч пар документов.

2.1. Метод на основе коллекции-медиатора

Как отмечено в [10], одним из наиболее эффективных методов кросс-языкового перехода является использование коллекции-медиатора (посредника), представляющей собой параллельный / псевдопараллельный корпус из документов на нескольких языках. Примерами таких корпусов могут выступать: мультиязычная энциклопе-

[1] https://digitallibrary.un.org/

дия Википедия[2], патентный массив (в сужении на патентные семейства, содержащие патенты на одно и то же изобретение, выданные в разных странах и описанные на национальных языках), цифровая библиотека ООН.

Явный семантический анализ. Рассмотрим для начала метод CL-ESA, описанный в [11]. В данной модели документы представляются в виде взвешенного вектора концептов, представленных статьями Википедии.

Мы отобрали около 800 тысяч пар выровненных статей на русском и английском языках в качестве медиатора концептов. В итоге, для каждого исследуемого документа *D*, вес концепта *C* определяется как косинус угла между вектором, состоящим из *M* наиболее значимых ключевых слов документа *D*, и вектором, состоящим из соответствующих ключевых слов для статьи, привязанной к концепту:

$$\frac{\sum_{w_i \in D} v_i c_i}{\sqrt{\sum_{w_i \in D} v_i^2}\sqrt{\sum_{w_i \in D} c_i^2}},$$

где $v_i - tf * idf$ мера слова w_i в документе D; $c_i - tf * idf$ мера слова w_i в привязанной к концепту статьи.

Для предварительного составления векторов концептов мы использовали топ из 200 ключевых слов (с весом более 0.05) для документа, чтобы вычислить веса концептов. Для каждого документа мы отобрали не более 1200 концептов с наибольшим весом. Поскольку мы строили вектора статей Википедии, используя сами статьи в качестве концептов, мы исключили из векторов концепты, отвечающие за те же статьи.

Упрощенная модель. Далее мы решили несколько упростить описанную выше модель: вместо построения векторного пространства концептов, решили использовать транзитивность функции сравнения документов. Иными словами, пусть у нас есть моноязычная функция поиска тематически похожих документов $F(d)$ и параллельный корпус документов $\{A_i, A_i^*\}$. Тогда, для того чтобы найти англоязычные документы, тематически похожие на заданный русскоязычный документ $d \in D_{rus}$, используется следующий алгоритм:

1. Ищем тематически похожие документы в коллекции-медиаторе A.
2. Отбираем соответствующие документы из медиатора A*.
3. Ищем похожие документы для найденных на Шаге 2.

Описанный подход позволяет использовать любую коллекцию-медиатор на уже существующем поисковом индексе, вместо того чтобы рассчитывать новое векторное пространство для каждого выбранного медиатора.

В качестве метода моноязычного поиска тематически похожих документов мы используем подход, описанный в [12], отлично подходящий для обработки больших массивов данных.

В качестве медиатора мы использовали как описанный выше набор из Википедии, там и нашу тестовую коллекцию из электронной библиотеки ООН: в ходе поиска мы исключали сравниваемые пары документов из медиатора.

2.2. Методы на основе векторных представлений слов

Принципы дистрибутивной семантики, впервые описанные в [13], играют особую роль при анализе текстов на естественном языке. Основная суть данных подходов заключается в построении такого векторного пространства, что семантически схожие слова будут иметь близкие вектора (например, по косинусной мере).

ETM. Развивая идею использования медиатора для перехода между двумя языками, в [14] был предложен метод, основанный на использовании эмбеддингов в качестве концептов, работающий для вероятностных поисковых моделей, например, с использованием известной функции ранжирования BM25[15]. Авторы расширяют меры tf и idf для слова t в документе d следующим образом:

$$\widehat{tf}_{t,d} = tf_{t,d} + \sum_{t' \in R(t)} P_T(t|t') tf_d(t');$$

$$\widehat{df}_t = |\{d \in D : t \in T_d \vee \exists t' \in R(t), t' \in T_d\}|,$$

где $P_T(t|t')$ – вероятность, что слово *t* будет переведено как *t'*, $R(t)$ – множество вариантов перевода слова *t*.

Таким образом, получаем следующие величины:

длина документа $\hat{L}_d = \sum_{t \in \hat{T}_d} \widehat{tf}_d(t)$;

средняя длина документа $\widehat{avgdl} = \frac{1}{|D|}\sum_{d \in D} \hat{L}_d$.

В итоге для BM25 получаем следующее выражение:

$$BM25_{ET}(q,d) = \\ = \sum_{t \in \hat{T}_d \cap T_q} \frac{(k_1+1)\widehat{tf}_d(t)}{k_1 + \widehat{tf}_d(t)} \frac{(k_2+1)tf_q(t)}{k_2 + tf_q(t)} log \frac{|D|+0.5}{\widehat{df}_t + 0.5},$$

[2] https://www.wikipedia.org

где $\widehat{\overline{tf_d(t)}} = \frac{\widehat{tf}_d(t)}{\hat{B}(d)}$, $\hat{B}(d) = (1-b) + b\frac{\hat{L}(d)}{\widehat{avgdl}}$.

Retrieval based approach. Другой интересный подход, описанный в [16], предполагает построение двуязычного векторного пространства. Основная идея заключается в формировании псевдо-двуязычного построения из двух моноязычных для обучения модели. Например, для предложений «мама мыла раму» и “mother washed a beautiful frame” получаем «мама mother мыла washed раму beautiful frame».

Нами была реализована данная модель в [17], а в качестве набора данных использовали набор параллельных предложений из следующих корпусов [18]:

- News Commentary;
- TED Talks 2013;
- MultiUN (first 2 million sentences);
- Wiki;
- JW300;
- QED;
- Tatoeba;
- Параллельный корпус Яндекса.

Предварительная обработка заключалась в разделении каждого предложения на токены, лемматизация токенов и парсинг текстов. Мы использовали AOT для русского языка и Udpipe для английского. Кроме того, убрали из предложений слова, относящиеся к служебным и семантически незначимым частям речи: союзы, местоимения, предлоги и т.п., а также общеупотребительные стоп-слова. После этого убрали пары, в которых разница по количеству слов превышала десять. Для формирования словаря использовались синтаксические конструкции длиной до четырех слов – были использованы только характерные для корпуса сочетания, т.е., встречающиеся более десяти раз. Например, для предложения, содержащего конструкцию “Russian presidential election …” мы построили 3 варианта с разными сочетаниями:

- “Russian_presidential_election …”;
- “Russian_election presidential_election …”;
- “Russian presidential election …”.

В итоге был сформирован корпус из более чем 5.1 миллиона предложений (более 10 миллионов – с учётом вариантов сочетаний). Полученный словарь содержал около 680000 слов/словосочетаний.

Для поиска была использована специальная реализация обратного индекса [12]: были проиндексированы слова и синтаксические конструкции (до четырех слов) с весами (TF*IDF), которые отражают соответствие слова документу:

$$TF_D(w_i) = \log_{len(D)+1}(Cnt(w_i) + 1);$$

для слова w_i из документа D

$$IDF = \max(0, \log_{10} \frac{N - w_{cnt} + 0.5}{w_{cnt} + 0.5}),$$

где N – общее количество документов в коллекции.

Итак, для поиска мы берем топ ключевых слов и словосочетаний из эталонного документа, соотносим их со словосочетаниями на другом языке, используя кросс-языковые эмбеддинги, а затем выделяем из индекса соответствующие документы, объединяя их во взвешенные вектора. После этого мы сравниваем целевые вектора с остальными по косинусной мере и расстоянию Хэмминга.

Сравнение тематик (top2vec). Также был рассмотрен подход, основанный на сравнении документов по тематикам: для каждого документа выделяются ключевые слова по tf*idf мере, для которых строятся эмбеддинги в двуязычном пространстве. В результате каждый документ представляется в виде усредненного вектора своих ключевых слов. Для сравнения векторов была использована косинусная мера.

LASER. В [19] была предложена универсальная многоязычная модель LASER (Language-Agnostic Sentence Representation). В данной системе используется пятислойный двунаправленный LSTM энкодер для построения эмбеддингов предложений в связке со вспомогательным декодером. Модель была обучена для 93 языков на открытых данных:

- Europarl: 21 европейский язык;
- United Nations: первые 2 миллиона предложений на арабском, российском и китайском;
- OpenSubtitles2018: параллельный корпус субтитров для фильмов на 57 языках;
- Global Voices: новостные заметки (38 языков);
- Tanzil: перевод Корана на 42 языка;
- Tatoeba.

В итоге предложение представляется в виде вектора размерности 1024 в едином мультиязычном пространстве.

Для поиска по векторному пространству используется метод приблизительных ближайших соседей, в качестве поискового индекса выступает Faiss IVFFlatIndex [20].

Мы использовали две схемы поиска документов:

- Сравнение предложений**.** В данном случае для каждого предложения каждого документа модель LASER строит вектор, который кладется в индекс. В ходе поиска мы ищем ближайшие по косинусной мере предложения для каждого предложения из заданного документа. Ранжирование поисковой выдачи происходит в соответствии с количеством подошедших предложений. В ходе индексации мы исключили предложения, содержащие меньше десяти символов – номера страниц, названия разделов и т.п.
- Сравнение по ключевым словам**.** Поскольку объем данных достаточно большой, а размерность векторного пространства модели велика, было принято решение испытать сравнение тематик, но в векторном пространстве модели LASER.

3. Экспериментальное исследование методов кросс-языкового поиска нормативно-правовых документов

Для сравнения качества работы описанных методов мы использовали стандартные метрики: точность (precision – P@1), полноту на N элементах (recall at N – Rec@N) и MAP на N элементах (mean average precision – MAP@N):

- Точность на K (precision at K) $P@K = \frac{\sum_{k=1}^{K} r^{true}(\pi^{-1}(k))}{K}$.
- Средняя точность на K (average precision at K) $ap@K = \frac{1}{K}\sum_{k=1}^{K} r^{true}\left(\pi^{-1}(k)\right) \cdot p@k.$
- Mean average precision at K $MAP@K = \frac{1}{N}\sum_{i=1}^{N} ap@K_{i.},$

где $r^{true}: E \to [0,1]$ – эталонная функция релевантности.

Для каждой пары нашего корпуса мы искали англоязычные документы, соответствующие русскоязычному. Результаты собраны в Табл. 1.

Используемые обозначения:

- CL-ESA – cross-lingual explicit semantic analysis;
- ETM – extended translation model;
- RBA – retrieval-based approach;
- LASER – Language-Agnostic Sentence Representations;
- Cos – косинусное расстояние;
- Ham – расстояние Хэмминга.

Отметим, что наилучшие результаты были достигнуты при использовании связки ETM и BM25. Также наглядно продемонстрировано, что в методах, основанных на построении медиатора, значительно более высокое качество достигается с применением тематически специализированной коллекции, нежели более общей.

4. Модификация методов кросс-языкового поиска нормативно-правовых документов с применением ссылочного анализа

Несмотря на вполне приемлемые результаты метода ETM+BM25 на фоне остальных, все они обладают слабой стороной: специфика юридических документов такова, что они сильно

Табл. 1. Метрики по размерным группам

Метод	P@1	Rec@5	Rec@10	Rec@20	Rec@150	MAP@150
CL-ESA	0.13	0.24	0.3	0.36	0.48	0.186
CL-ESA, упрощ., Википедия	0.0016	0.004	0.0075	0.01	0.036	0.0037
CL-ESA, упрощ, ООН	0.19	0.29	0.365	0.427	0.535	0.246
ETM+BM25	0.78	0.87	0.89	0.9	0.91	0.82
RBA, cos	0.58	0.73	0.77	0.8	0.83	0.645
RBA, ham	0.62	0.75	0.79	0.81	0.83	0.678
Top2Vec	0.7	0.81	0.83	0.85	0.89	0.749
LASER, тематики	0.05	0.118	0.15	0.19	0.33	0.088
LASER, по предложениям	0.05	0.27	0.415	0.55	0.767	0.16

Табл.2. Метрики качества выделения ссылок для русскоязычных документов

Модель	**F_1**	**P**	**R**
SBERT_ BASELINE	0.81	0.81	0.80
SIAM_SBERT_FEED	0.83	0.83	0.83
SIAM_SBERT_ TR	0.85	0.85	0.85

Табл. 3. Метрики качества выделения ссылок для документов на русском и английском языках

Модель	**F_1**	**P**	**R**
SBERT_ BASELINE	0.79	0.80	0.78
SIAM_SBERT_FEED	0.56	0.59	0.58
SIAM_SBERT_ TR	0.54	0.56	0.57

фрагментированы и содержат сразу множество тематик, что может приводить к пропускам релевантных документов. В [21] рассмотрен принципиально другой подход к поиску релевантных документов, а именно – анализ неявных ссылок между документами.

Предложенный метод основан на использовании Сиамских сетей с использованием модели эмбеддингов SentenceBERT [22]. Документы разбиваются на фрагменты по 30 токенов, для которых строятся фрагментные эмбеддинги, в последствии объединяемые в эмбеддинги документов. Считается, что между документами есть связь, если хотя бы один фрагмент из первого документа соотносится с хотя бы одним фрагментом из другого.

Однако при использовании данного подхода на двуязычном параллельном корпусе наблюдается значительное падение качества в сравнении с моноязычным корпусом (Табл. 2 и 3). Для оптимальной настройки мультиязычной сети, использующей модель SBERT, необходим размеченный (по ссылкам) параллельный (по предложениям) корпус. Мы полагаем, что использование ссылочного анализа позволит расширить полноту при проведении глубокого анализа и информационного поиска в области юридических документов в сравнении с использованием только методов поиска тематически похожих документов.

Заключение

В работе проведено сравнение различных подходов и моделей к решению задачи кроссязыкового поиска тематически похожих правовых документов, в результате чего мы приходим к выводу, что одним из лучших решений будет использование модели ETM с функцией ранжирования BM25.

По результатам проведенного исследования на базе технологии TextAppliance[3] был построен экспериментальный стенд для проведения аналитического поиска юридических документов. Стенд позволяет проводить семантический и эксплоративный поиск как на языке заданного эталона, так и кросс-языковой, выявлять текстовые заимствования и цитирования между документами, автоматически строить краткий реферат документа, выделять ключевую лексику и многое другое, что делает его крайне полезным инструментом при проведении аналитических исследований нормативно-правовой документации.

Литература

1. Dini L., Peters W., Liebwald D., Schweighofer E., Mommers L.,Voermans W. Cross-lingual legal information retrieval using a WordNet architecture, in Proceedings of the 10th international conference on Artificial intelligence and law. Bologna. Italy. 2005.
2. Абрамова Н. Н., Глобус Е. И. Формирование многоязычных словарей и их использование при кросс-языковом поиске информации. Интернет-математика. Автоматическая обработка веб-данных. 2005. С. 18-37.
3. Curtoni P., Dini L., Tomaso V. D., Mommers L., Peters W., Quaresma P., Schweighofer E., Tiscornia D. Semantic access to multilingual legal information. 1999.
4. Oard D.W., Hackett P. Document translation for cross-language text retrieval at the University of Maryland. The 6th Text Retrieval Convference (TREC-6). E.M. Voorchees and D.K. Harman. 1998.
5. McCarley J.S. Should we translate the documents or the queries in cross-language information retrieval? ACL'99: Proceedings of the 37 annual meeting of the Association

[3] http://www.textapp.ru

for Computational Linguistics on Computational Linguistics. 1999. P. 208-214.
6. Dumais S., Letsche T., Littman M., Landauer T. Automatic cross-language retrieval using latent semantic indexing. AAAI Spring Symposium on Cross-Language Text and Speech Retrieval. 1997. P. 18-24.
7. Chandar A.P.S., Lauly S., Larochelle H., Khapra M., Ravindran B., Raykar V.C., SahaA. An autoencoder approach to learning bilingual word representations. Proc. 27th International Conference on Neural Information Processing Systems. 2014. P. 1853-1861.
8. Mueller J., Thyagarajan A. Siamese recurrent architectures for learning sentence similarity. Proc. 30th AAAI Conference on artificial intelligical intelligence. 2016. P. 2786-2792.
9. Seki K. On cross-lingual text similarity using neural translation models. Journal of Information Processing. Vol. 27. 2019. P. 315-321.
10. Жебель В.В., Крескин А.Д., Соченков И.В.: Кросс-языковой анализ юридических документов. Труды ИСА РАН. 2020. Т.70. №1. С.24-29.
11. Potthast M., Barrón-Cedeño A., Stein B., Rosso P. Cross-language plagiarism detection. Language Resources and Evaluation. 2011. №45(1). P. 45–62.
12. Sochenkov I.V., Zubarev D.V., Tikhomirov I.A. Exploratory patent search. Informatics and its Applications.2018. №12 (1). P. 89-94.
13. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. In: ICLR Workshop. 2013.
14. Rekabsaz N., Lupu M., Hanbury A., Zuccon G. Generalizing translation models in the probabilistic relevance framework. In: Proceedings of CIKM. 2016.
15. Robertson S.E. et al. Okapi at TREC-3.0. In: Proceedings of the Third Text REtrieval Conference (TREC 1994). Gaithersburg USA. November.1994.
16. Vulić I., Moens M.F. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015). Vol. 2. P.719–725.
17. Zubarev D.V., Sochenkov I.V. Cross-lingual similar document retrieval methods. Proceedings of the Institute for System Programming. 2019. №31 (5). P.127–136.
18. Tiedemann J. Parallel Data, Tools and Interfaces in OPUS. In: Proc. of the language resources and evaluation (LREC). 2012. P.2214-2218.
19. Artetxe M., Schwenk H. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. Transactions of the Association for Computational Linguistics. 2019). №7. P.597–610.
20. Johnson J., Douze M., Jégou H. Billion-scale similarity search with GPUs. arXiv:1702.0873. 2017.
21. Devyatkin D., Pogorelskaya Y., Yadrintsev V., Sochenkov I. Detection of Missed Links in Large Legal Corpora. 2021 Ivannikov Memorial Workshop (IVMEM). 2021. P. 23-27.
22. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2019. P.3982–3992.

Жебель Владимир Викторович. Научный сотрудник. Общество с ограниченной ответственностью «Технологии системного анализа». Области исследований: искусственный интеллект, обработка больших массивов данных, компьютерная лингвистика. E-mail: zhebel@isa.ru

Девяткин Дмитрий Алексеевич. Научный сотрудник. Федеральный исследовательский центр «Информатика и управление» Российской академии наук. Области исследований: извлечение информации из текстов, обработка больших массивов данных, компьютерная лингвистика, наукометрия и научно-техническое прогнозирование. E-mail: devyatkin@isa.ru

Зубарев Денис Владимирович. Младший научный сотрудник. Федеральный исследовательский центр «Информатика и управление» Российской Академии Наук. Области исследований: искусственный интеллект, информационный поиск, поиск текстовых заимствований. E-mail: zubarev@isa.ru

Соченков Илья Владимирович. Кандидат физико-математических наук. Ведущий эксперт-консультант. Университет Иннополис, Казань. Ведущий научный сотрудник, Институт системного программирования им. В.П. Иванникова Российской академии наук. Техник 1-й категории, Федеральный исследовательский центр «Информатика и управление» Российской академии наук. Области исследований: интеллектуальные методы поиска и анализа информации, обработка больших массивов данных, контентная фильтрация, компьютерная лингвистика, распознавание образов. E-mail: sochenkov@isa.ru

Methods for Cross-Lingual Retrieval of Similar Documents in Legal Domain Based on Machine Learning

V. V. Zhebel[I], D. A. Devyatkin[II], D. V. Zubarev [II], I. V. Sochenkov [II,III,IV]

[I]Limited liability company «Technologies for systems analysis», Moscow, Russia
[II]Federal Research Center «Computer Science and Control» of the Russian Academy of Sciences, Moscow, Russia.
[III]Innopolis University, Kazan, Russia
[IV]Ivannikov Institute for System Programming of the Russian Academy of Sciences, Moscow, Russia

Abstract. The need of studying the international experience to improve legislation cause the need of information retrieval systems to be good in multilingual legal domain. One of the possible solutions is thematically similar document retrieval. However, there is an important task to transfer between languages to let the user put a document on the one language and get the search result on another one. The paper describes different approaches to solve this problem: from classical mediator-based methods to modern procedures of distributive semantics. As a test collection, we have used the UN digital library. The combination of the extended translation model and BM25 ranking function demonstrates the best results.

Keywords: Cross-Lingual Document Retrieval, Distributional Semantics, Information Retrieval in the Legal Domain.

DOI 10.14357/20718594220203

References

1. Dini L., Peters W., Liebwald D., Schweighofer E., Mommers L., Voermans W. Cross-lingual legal information retrieval using a WordNet architecture," in Proceedings of the 10th international conference on Artificial intelligence and law. Bologna, Italy. 2005.
2. Abramova N.N. , Globus E.I. Formation of multilingual dictionaries and their use in cross-language information retrieval. pp. 18-37, 2005. P. Curtoni, L. Dini, V. D. Tomaso, L. Mommers, W. Peters, P. Quaresma, E. Schweighofer and D. Tiscornia, Semantic access to multilingual legal information.1999.
3. Curtoni P., Dini L., Tomaso V. D., Mommers L., Peters W., Quaresma P., Schweighofer E., Tiscornia D. Semantic access to multilingual legal information. 1999.
4. Oard D.W., Hackett P. Document translation for cross-language text retrieval at the University of Maryland. The 6th Text Retrieval Convference (TREC-6). E.M. Voorchees and D.K. Harman. 1998.
5. McCarley J.S. Should we translate the documents or the queries in cross-language information retrieval? ACL'99: Proceedings of the 37 annual meeting of the Association for Computational Linguistics on Computational Linguistics. 1999. P. 208-214.
6. Dumais S., Letsche T., Littman M., Landauer T. Automatic cross-language retrieval using latent semantic indexing. AAAI Spring Symposium on Cross-Language Text and Speech Retrieval. 1997. P. 18-24.
7. Chandar A.P.S., Lauly S., Larochelle H., Khapra M., Ravindran B., Raykar V.C., SahaA. An autoencoder approach to learning bilingual word representations. Proc. 27th International Conference on Neural Information Processing Systems. 2014. P.1853-1861.
8. Mueller J., Thyagarajan A. Siamese recurrent architectures for learning sentence similarity. Proc. 30th AAAI Conference on artificial intelligical intelligence. 2016. P.2786-2792.
9. Seki K. On cross-lingual text similarity using neural translation models. Journal of Information Processing. Vol. 27. 2019. P.315-321.
10. Zhebel, V., Kreskin, A., Sochenkov, I.: Cross-lingual document analysis in legal domain. Trudy Instituta sistemnogo analiza rossiyskoy akademii nauk. 2020. 70(1). P. 24–29.
11. Potthast M., Barrón-Cedeño A., Stein B., Rosso P. Cross-language plagiarism detection. Language Resources and Evaluation .2011.45(1). P.45–62.
12. Sochenkov I.V., Zubarev D.V., Tikhomirov I.A. Exploratory patent search. Informatics and its Applications.2018. 12 (1). P. 89-94.
13. Mikolov, T., Chen, K., Corrado G., and Dean, J.: Efficient estimation of word representations in vector space. In: ICLR Workshop. 2013.
14. Rekabsaz N., Lupu M., Hanbury A., Zuccon G. Generalizing translation models in the probabilistic relevance framework. In: Proceedings of CIKM. 2016.
15. Robertson S.E. et al. Okapi at TREC-3.0. In: Proceedings of the Third Text REtrieval Conference (TREC 1994). Gaithersburg, USA, November. 1994.
16. Vulić I., Moens M.F. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015. Vol. 2. P.719–725.
17. Zubarev D.V., Sochenkov I.V. Cross-lingual similar document retrieval methods. Proceedings of the Institute for System Programming. 2019. 31 (5). P.127–136.

18. Tiedemann J. Parallel Data, Tools and Interfaces in OPUS. In: Proc. of the language resources and evaluation (LREC). 2012. P.2214-2218.
19. Artetxe M., Schwenk H. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. Transactions of the Association for Computational Linguistics. 2019.7. P.597–610.
20. Johnson J., Douze M., Jégou H. Billion-scale similarity search with GPUs. arXiv:1702.08734. 2017.
21. Devyatkin D., Pogorelskaya Y., Yadrintsev V., Sochenkov I. Detection of Missed Links in Large Legal Corpora. 2021 Ivannikov Memorial Workshop (IVMEM). 2021. P.23-27.
22. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2019, P.3982–3992.

Zhebel Vladimir V. Research fellow, limited liability company “Technologies for systems analysis”. Research areas: artificial intelligence, big data, natural language processing. E-mail: zhebel@isa.ru

Devyatkin Dmitry A. Research fellow, Federal Research Center “Computer Science and Control”, the Russian Academy of Sciences. Research areas: text information retrieval, big data, natural language processing, scientometrics, scientific and technical foresight. E-mail: devyatkin@isa.ru

Zubarev Denis V. Junior research fellow, Federal Research Center “Computer Science and Control”, the Russian Academy of Sciences. Research areas: artificial intelligence, information retrieval, plagiarism detection. E-mail: zubarev@isa.ru

Sochenkov Ilya V. Candidate of physical and mathematical sciences. Leading Expert Consultant, Innopolis University. Lead Research Fellow, Ivannikov Institute for System Programming of the Russian Academy of Sciences. Junior research technician, Federal Research Center “Computer Science and Control”, the Russian Academy of Sciences. Research areas: intellectual methods for information search and analysis, big data, content filtrering, natural language processing, pattern recognition. E-mail: sochenkov@isa.ru