Кросс-языковой анализ юридических документов*

В.В. Жебель[I], А.Д. Крескин[II], И.В. Соченков[I]

[I] Федеральный исследовательский центр «Информатика и управление» Российской академии наук
[II] ООО «Технологии Системного Анализа»

Аннотация. В настоящее время особую актуальность приобретает сравнительный правовой анализ международного законодательства по вопросам права цифровых технологий. При проведении такого анализа значительную поддержку может оказывать возможность автоматизированного кросс-языкового поиска и выделения эквивалентных формулировок. В данной статье рассмотрены основные существующие методы кросс-языкового анализа документов, а также вопрос их применимости к юридическим документам.

Ключевые слова: *кросс-языковой поиск, обработка естественного языка, правовые исследования.*
DOI: 10.14357/20790279200103

Введение

В наши дни объемы существующей юридической информации настолько велики, что свободная навигация по ней и поиск интересующей информации без использования специальных инструментов даже на русском языке является сложной задачей. В случае сопоставления правовых документов разных стран возникает ещё большая проблема, связанная с тем, что юридические термины в иностранных языках в значительной степени отличаются между собой, что требует от исследователя высокого уровня владения сразу несколькими языками. Для облегчения задачи поиска необходимой информации применяются специальные компьютерные средства автоматизации. В данной работе приводится краткий обзор актуальных на сегодняшний день систем кросс-языкового поиска информации.

Основной целью данной статьи является исследование вопроса перехода между двумя языками, а не рассмотрение полных алгоритмов поиска и выделения информации.

1. Кросс-языковой поиск

В классических задачах поиска и извлечения информации и запрос, и документы, среди которых проводится поиск, написаны на одном и том же языке. Для поиска документов, написанных на языке, отличном от языка запроса, следует использовать дополнительную процедуру преобразования, которая обеспечила бы связь между разными языками, переводя сам запрос на целевой язык или документы (в некоторых случаях – их предварительно построенные аннотации) на исходный язык запроса.

На данный момент можно выделить две основные большие группы подходов к вопросу перевода:

– Системы с заранее подготовленным экспертами набором тезаурусов и онтологий [1] [2].
– Системы, основанные на машинном обучении [3] [4] [5] [6].

Рассмотрим далее особенности представителей этих двух групп.

1.1. Методы на основе многоязычных словарей

Наиболее очевидным решением проблемы автоматического перевода является идея использования того же инструмента, что используют люди-переводчики: заранее подготовленные двуязычные словари. Следует отметить, что, поскольку, точный перевод слов будет зависеть от контекста и области, к которой принадлежит текст, то, вообще говоря, требуются подобрать предметные тезаурусы и онтологии.

Вообще говоря, существует большое количество систем машинного перевода, использующих тезаурусы и онтологии: Lingvo, MultiLex, Eckado, RetrievalWare WordNet [7] и многие другие. Одним из примеров использования можно назвать систе-

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

му, внедрённую в ВИНИТИ [8] для поиска в реферативных русскоязычных базах по запросу на английском языке, в которой для перевода запросов используется ERTRANS.

Отдельный интерес представляет процесс формирования таких тезаурусов. Так, например, в исследовании, описанном в [2], приводится описание автоматизированного формирования многоязычных словарей на основе параллельных данных из сети. В этой же работе был описан и протестирован метод перевода пользовательских запросов на другой язык. Для перевода авторы проанализировали протоколы работы поисковой системы «Яндекс», из которой были извлечены реальные запросы пользователей. Запросами, направленными на поиск иноязычной информации, авторы обозначили те, в которых встречаются слова на иностранном языке.

Формализация и перевод происходил в несколько этапов. Сначала были проведены стандартные процедуры очистки запроса от знаков препинания, производится морфологический анализ текста запроса, из него выделяются наименования понятий, эти понятия ищутся в многоязычном словаре, после чего формируется иноязычный запрос, составленный из переведенных понятий и частей запроса, не требующих перевода.

По итогам тестирования с использованием словаря в 20000 лексических единиц удалось перевести примерно 75% запросов, что является хорошим результатом, с учетом автоматизированного построения словаря.

Стоит сказать, что для подготовки методов, применение которых предполагается в специализированных областях, лучше использовать данные, связанные с заданной областью. Однако ручная очистка большого количества данных может занять много времени. Один из вариантов решения данной проблемы – автоматизированная классификация данных.

В работе [9] авторы описали способ классификации многоязычного корпуса документов на основе зависимостей переводов слов документов. Классифицированные по темам данные можно использовать для обучения моделей для конкретной предметной области.

1.2. Использование параллельных корпусов

Далее рассмотрим применение методов машинного обучения с использованием параллельных корпусов для автоматического перевода. Основным способом обучения моделей в данном случае является использование параллельных корпусов текстов на двух языках, что позволяет сформулировать отношения между словами из разных языков.

Часто в качестве параллельных корпусов используются статьи на разных языках из Википедии. Так, например, в [3] было наглядно продемонстрировано использование Википедии для обучения модели LDA (Latent Dirichlet allocation). В работе [10] также используется модель кросс-языкового поиска похожих документов, обученная на документах из Википедии. Следует отметить, что поскольку корпуса таких статей не только не являются выровненными, но и различаются по масштабам (если в русскоязычном сегменте Википедии представлено порядка 1.5 миллионов статей, то в англоязычном – около 6 миллионов), что делает её далеко не идеальной обучающей выборкой, несмотря на распространённость.

Не менее интересный пример представлен в [4], где исследовались возможности многоязычного поиска информации на основе вероятностной модели и построения параллельных корпусов текстов, автоматически созданных на основе информации из сети за счёт использования версий сайтов на разных языках. Кроме того, авторы провели сравнение рассматриваемых моделей с системами двуязычных словарей, и словари проигрывают по эффективности.

Основная идея метода перевода запроса в выборе из него тех слов, которые с большой вероятностью встретятся в любом из переводов запроса. Это сделано для повышения качества перевода, путем избавления от «шумов», слов с низкой встречаемостью в корпусах, а также для увеличения скорости обработки запроса.

В результате тестирования средняя точность «перевода» запроса составила около 70% на модели, обученной на веб-данных и около 80% на модели, обученной на корпусе Hansard.

Другими исследователями, решившими проблему похожим способом, стали авторы работы [11]. В своем исследовании ими было разработано расширение для стандартной модели LDA, поддерживающее двуязычные данные, названное ими BiLDA. Эта модель использовалась для кросс-языкового поиска и тестировалась на коллекциях CLEF. Результаты авторы оценивали при помощи меры MRR, равной для их экспериментов с парой языков Английский-Голландский 0.3506, что близко к средним показателям работающих систем на тот момент.

Рассмотрим также возможность перевода запросов пользователей при помощи дискриминационной модели. Такой метод в 2011 году запатен-

товала Microsoft Corporation [12]. Идея метода в обучении модели на предварительно подготовленных данных, содержащих как моноязычные, так и мультиязычные параллельные тексты.

Система также способна анализировать запросы пользователей, чтобы принять решение, хочет ли пользователь получить данные только на языке запроса, или же ему нужен более широкий спектр ответов. В случае принятия решения о поиске на ином языке, система подбирает возможные языки для перевода на основе внутреннего ранжирования, а затем при помощи модели, обученной на параллельных корпусах, производит перевод.

В дальнейшем в 2014 году Microsoft также получила патент на систему мультиязычного поиска [13], предназначенную для определения ссылок на ресурсы, язык которых отличен от языка запроса.

1.3. Обучение с использованием моноязычных данных

Особого внимания следует удостоить метод, разработанный учеными Мангеймского университета в сотрудничестве с исследователями из Оксфорда [5], который интересен в отношении рассматриваемой проблемы тем, что для начального обучения модели достаточно использовать моноязычные корпуса документов.

Метод основывается на сопоставлении векторных представлений взаимосвязей слов в контекстах разных языков. При разработке использовалось построение моделей на основе машинного обучения без учителя на основе больших корпусов текстов (Wikipedias), или предварительно вычисленные модели FastText [14] [15] для пар языков Английский-Датский, Английский-Итальянский и Английский-Финский.

Для сравнения и тестирования были приведены две модели, обученные на параллельных корпусах документов и на параллельных корпусах терминов. Для представления терминов запроса и документа в понятной для моделей формы применялись взвешенная и невзвешенная сумма векторов, а также прямой перевод терминов с помощью словаря.

Тестирование производилось на тестовых коллекциях CLEF 2000-2003 годов. Результаты тестирования и сравнительные характеристики для всех моделей и методов представления терминов запроса можно видеть на табл. 1 [5]. Для оценки авторами использована мера MAP (mean Average Precision).

Отметим, что предложенный метод обучения без учителя на моноязычных корпусах показывает значительно лучшие результаты.

Табл. 1.

MAP для алгоритмов CLIR основанных на разных моделях

CL Embs	Model	EN→NL			EN→IT			EN→FI	
		2001	2002	2003	2001	2002	2003	2002	2003
-	LM-UNI	.119	.196	.136	.085	.167	.137	.111	.142
CL-CD	BWE-Agg-Add	.111	.138	.137	.087	.114	.147	.026	.084
	BWE-Agg-IDF	.144	.203	.189	.127	.157	.188	.082	.125
	TbT-QT	.125	.196	.120	.106	.148	.143	.176	.140
	Линейная комбинация BWE-Agg-IDF и TbT-QT (λ=0.5)	.145	.216	.174	.120	.183	.216	.179	.189
	Линейная комбинация BWE-Agg-IDF и TbT-QT (λ=0.7)	.142	.216	.180	.127	.180	.207	.183	.197
CL-WT	BWE-Agg-Add	.149	.168	.203	.138	.155	.236	.078	.217
	BWE-Agg-IDF	.185	.196	.243	.169	.166	.248	.086	.204
	TbT-QT	.159	.164	.176	.129	.150	.218	.095	.095
	Линейная комбинация BWE-Agg-IDF и TbT-QT (λ=0.5)	.202	.198	.280	.187	.168	.228	.117	.190
	Линейная комбинация BWE-Agg-IDF и TbT-QT (λ=0.7)	.202	.198	.263	.181	.171	.230	.120	.164
CL-UN-SUP	BWE-Agg-Add	.125	.153	.198	.119	.126	.213	.078	.239
	BWE-Agg-IDF	.172	.204	.250	.157	.161	.253	.102	.223
	TbT-QT	.229	.257	.299	.232	.257	.345	.145	.243
	Линейная комбинация BWE-Agg-IDF и TbT-QT (λ=0.5)	.258	.300	.330	.225	.248	.325	.154	.307
	Линейная комбинация BWE-Agg-IDF и TbT-QT (λ=0.7)	.259	.303	.336	.236	.253	.347	.151	.307

2. Применимость в юридической сфере

Возвращаясь к первоначальному вопросу, проанализируем возможности использования рассмотренных методов применительно к правовым документам. Хоть юридические документы и представляют из себя тексты на естественном языке, они обладают определённой спецификой в части используемой лексики и обладают специфической структурой оформления.

Так, использование методов, основанных на двуязычных словарях, потребует значительных усилий в подготовке предметных тезаурусов и онтологий в юридической сфере, а результаты, как показано ранее, едва ли превзойдут возможности методов, основанных на машинном обучении. Например, вариант системы кросс-языкового поиска юридической информации, основанной на параллельных корпусах терминов для швейцарского и английского языков, был описан в работе [16]: построены словари юридических терминов для швейцарского и английского языков, использовался прямой перевод терминов. В результатах тестирования были получены результаты Average Precision около 33%, что в 2006 году было показателем выше среднего.

Методы, использующие для создания модели машинного обучения параллельные корпуса, вполне могут быть использованы в правовой сфера благодаря значительному количеству международных документов, переведенных на несколько языков. Так, например, в свободном доступе присутствует массив резолюций ООН сразу на 6 мировых языках (английском, арабском, испанском, китайском, русском и французском) – хороший пример выровненного параллельного корпуса.

В случае же с методом, описанным в [5], ситуация складывается наилучшим образом: для каждого языка можно собрать огромные массивы нормативно-правовой документации, которые можно использовать для построения систем эмбеддингов, однако здесь необходим особый подход к вычислению линейного отображения между полученными пространствами.

Заключение

Как было показано в статье, наибольшей перспективностью обладают методы, основанные на машинном обучении. Так как русский является одним из «мировых» языков, то для него существуют достаточно большие объемы параллельных юридических документов. Более того, с учётом специфики исследуемой области – международного права – с большой вероятностью, лидирующие позиции будет занимать независимое построение векторных пространств на моноязычных данных с дальнейшим формированием отображения между ними.

Конечно, методы машинного обучения требуют большого количества данных, желательно предварительно подготовленных. В связи с этим для языков, использующихся мало или имеющих недостаточно информации для обучения моделей, можно использовать менее эффективные способы поиска, основанные на предварительно подготовленных словарях.

Дальнейшим направлением исследования является проверка обозначенной гипотезы, а также оценка эффективности такого подхода в рассматриваемой сфере. Для этого необходимо подготовить моноязычные массивы для обучения, а также параллельные двуязычные корпуса для сравнения с другими методами машинного обучения.

В случае успешных результатов, сформированные модели могут стать эффективным инструментом информационной поддержки сравнительного правового анализа в международном праве.

Литература

1. *L. Dini, W. Peters, D. Liebwald, E. Schweighofer*, L. Mommers and W. Voermans, "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. и Глобус Е.И.* "Формирование многоязычных словарей и их использование при кросс-языковом поиске информации," Интернет-математика 2005. Автоматическая обработка веб-данных, pp. 18-37, 2005.
3. *Roth, Benjamin & Klakow, Dietrich*. Combining Wikipedia-Based Concept Models for Cross-Language Retrieval. Lecture Notes in Computer Science. 2010. - сс.47-59..
4. *J.-Y. Nie, M. Simard, P. Isabelle and R. Durand*, "Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web," in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, California, USA, 1999.
5. *R. Litschko, S. P. Ponzetto, G. Glavaš and I. Vulić*, "Unsupervised Cross-Lingual Information Retrieval Using Monolingual Data Only," in SIGIR`18, Ann Arbor, Michigan, USA, 2018.
6. *J. Landthaler, B. Waltl, P. Holl and F. Matthes*, "Extending Full Text Search for Legal Document Collections using Word Embeddings," in

Proceedings of Jurix: International Conference on Legal Knowledge and Information Systems, Sofia Antopolis, France, 2016.
7. *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.
8. *Белоногов Г.Г. и др.* Компьютерная лингвистика и перспективные информационные технологии - М.: Рус. мир, 2004. - 248с..
9. *P. Prettenhofer and B. Stein,* "Cross-language text classification using structural correspondence learning," in Proceedings of the 48th annual meeting of the association for computational linguistics, 2010.
10. *M. Potthast, B. Stein and M. Anderka*, "A Wikipedia-Based Multilingual Retrieval Model," in European Conference on Information Retrieval, 2008.
11. *I. Vulić, W. D. Smet and M.-F. Moens*, "Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora," Information Retrieval, pp. 331-368, 2013.
12. *C. Niu and Ming Zhou*, "Cross-Lingual Query Suggestion". United States Patent US 8051061 B2, 1 11 2011.
13. *A.v. Kurochkin, A. Kamel And S.K. Parameswar*, "Multi-Language Information Retrieval and Advertising". United States Patent US 2014/0280295 A1, 18 09 2014.
14. *T. Mikolov, P. Bojanowski, E. Grave and A. Joulin*, Bag of Tricks for Efficient Text Classification, Facebook AI Research, 2016.
15. *T. Mikolov, A. Joulin, E. Grave and P. Bojanowski*, Enriching Word Vectors with Subword Information, Facebook AI Research, 2017.
16. *P. Sheridan, M. Braschlert and P. Schäuble*, "Cross-language information retrieval in a Multilingual Legal Domain," in Research and Advanced Technology for Digital Libraries, 2006, pp. 253-268.

Жебель Владимир Викторович Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), г. Москва. Младший научный сотрудник. Количество печатных работ: 12. Область научных интересов: интеллектуальные методы поиска и анализа информации, обработка больших массивов данных, компьютерная лингвистика. E-mail: zhebel@isa.ru

Соченков Илья Владимирович Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), г. Москва. Заведующий отделом «Интеллектуальные технологии и системы». Кандидат физико-математических наук. Количество печатных работ: 70. Область научных интересов: интеллектуальные методы поиска и анализа информации, обработка больших массивов данных, контентная фильтрация, компьютерная лингвистика, распознавание образов. E-mail: sochenkov@isa.ru

Крескин Алексей Дмитриевич ООО «Технологии Системного Анализа» (ООО «ТСА»), г. Москва. Программист научно-исследовательского отдела. Область научных интересов: интеллектуальные методы поиска и анализа информации, обработка больших массивов данных. E-mail: kreskin@tesyan.ru

Cross-language legal documents analysis

V.V. Zhebel[I], A.D. Kreskin[II], I.V. Sochenkov[I]

[I] Federal Research Center “Computer Science & Control” of the Russian Academy of Sciences (FRC CS&C RAS), Moscow, Russia.

[II] OOO “Technologies of System Analysis”, Moscow, Russia

Abstract: Nowadays the problem of the comparative analysis for the international legal domain of digital technologies is very important. In this case significant support could be provided by methods of automated cross-language information retrieval and extracting similar terms. This paper discusses the main approaches of cross-language analysis and their application to legal domain.

Keywords: *cross-language search, natural language processing, legal researches*.
DOI: 10.14357/20790279200103

References

1. *L. Dini, W. Peters, D. Liebwald, E. Schweighofer*, L. Mommers and W. Voermans, "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. and Globus E.I., " Formation of multilingual dictionaries and their use in cross-language information retrieval," Internet-matematika 2005. Avtomaticheskaya obrabotka veb-dannyh, pp. 18-37, 2005*
3. *Roth, Benjamin & Klakow, Dietrich*. Combining Wikipedia-Based Concept Models for Cross-Language Retrieval. Lecture Notes in Computer Science. 2010. - cc.47-59..
4. *J.-Y. Nie, M. Simard, P. Isabelle and R. Durand*, "Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web," in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, California, USA, 1999.
5. *R. Litschko, S. P. Ponzetto, G. Glavaš and I. Vulić*, "Unsupervised Cross-Lingual Information Retrieval Using Monolingual Data Only," in SIGIR`18, Ann Arbor, Michigan, USA, 2018.
6. *J. Landthaler, B. Waltl, P. Holl and F. Matthes*, "Extending Full Text Search for Legal Document Collections using Word Embeddings," in Proceedings of Jurix: International Conference on Legal Knowledge and Information Systems, Sofia Antopolis, France, 2016.
7. *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.
8. *Belonogov G.G. and others, Computational Linguistics and Advanced Information Technologies - M.: Rus. world, 2004. – 248p.*.
9. *P. Prettenhofer and B. Stein,* "Cross-language text classification using structural correspondence learning," in Proceedings of the 48th annual meeting of the association for computational linguistics, 2010.
10. *M. Potthast, B. Stein and M. Anderka*, "A Wikipedia-Based Multilingual Retrieval Model," in European Conference on Information Retrieval, 2008.
11. *I. Vulić, W. D. Smet and M.-F. Moens*, "Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora," Information Retrieval, pp. 331-368, 2013.
12. *C. Niu and Ming Zhou*, "Cross-Lingual Query Suggestion". United States Patent US 8051061 B2, 1 11 2011.
13. *A.v. Kurochkin, A. Kamel And S.K. Parameswar*, "Multi-Language Information Retrieval and Advertising". United States Patent US 2014/0280295 A1, 18 09 2014.
14. *T. Mikolov, P. Bojanowski, E. Grave and A. Joulin*, Bag of Tricks for Efficient Text Classification, Facebook AI Research, 2016.
15. *T. Mikolov, A. Joulin, E. Grave and P. Bojanowski*, Enriching Word Vectors with Subword Information, Facebook AI Research, 2017.
16. *P. Sheridan, M. Braschlert and P. Schäuble*, "Cross-language information retrieval in a Multilingual Legal Domain," in Research and Advanced Technology for Digital Libraries, 2006, pp. 253-268.

V.V. Zhebel. Federal Research Center “Computer Science & Control” of the Russian Academy of Sciences (FRC CS&C RAS), Moscow, Russia. Junior research scientist. Graduated from Lomonosov Moscow State University in 2013. Author of 12 scientific papers. Research interests: Information Retrieval, Machine Learning, Computational Linguistics. E-mail: zhebel@isa.ru

I.V. Sochenkov. Federal Research Center “Computer Science & Control” of the Russian Academy of Sciences (FRC CS&C RAS), Moscow, Russia. Head of the Department for Intelligent Technologies & Systems. He graduated from the Peoples’ Friendship University of Russia (RUDN University) in 2009. PhD in Computer Science. Author of 70 scientific papers. Research interests: Information Retrieval Machine Learning, Content Filtering, Computational Linguistics, Pattern Recognition. E-mail: sochenkov@isa.ru

A.D. Kreskin. OOO “Technologies of System Analysis”, Moscow, Russia. Research programmer. Graduated from Peoples’ Friendship University of Russia (RUDN University) in 2018. Research interests: information retrieval machine learning, Big Data analysis. E-mail: kreskin@tesyan.ru