чениях параметров): *recall* = 0,84 и *precision* = 0,46. Это показало бóльшую эффективность разработанного модуля, при этом в отличие от аналогичных нейросетевых [6], он является более универсальным и может использоваться для локализации произвольных текстовых объектов (строк, слов и др.). Также получена оценка в рамках решения задачи локализации слов 233 тестовых изображений базы ICDAR 2013 (см. рис. 5г), которая выражалась тремя параметрами: $rec = N/K$, $prec = N/P$ (где N – число верно локализованных слов, K – общее число слов, P – общее число локализованных слов) и $F - score = 2rec \cdot prec/(rec + prec)$, при расчете которых применяется специализированная система штрафов в ситуациях соответствия «один ко многим» и «многие ко многим». Оценка модуля составила: *rec* = 78,71, *prec* = 71,03, *F–score* = 74,67, коммерческого аналога ABBY OCR SDK v10 – 35,07, 60,95 и 44,52, а нейросети со значительно более громоздкой архитектурой – 89,53, 94,26 и 91,84 соответственно [13]. Данная оценка показывает работоспособность предложенной модели детектора в условиях сложной композиции изображений и высокой стилистической вариативности текста. Возможности повышения точности локализации связаны с увеличением количества рассматриваемых масштабов изображения и совершенствованием процедуры сегментации блоков на слова.

Заключение. В исследовании показана перспективность применения компактных СНС для решения практических задач обработки текста на изображениях. Преимуществом компактных СНС, по сравнению с ГНС, является достаточно высокая обобщающая способность в сочетании с возможностью их реализации на неспециализированных языках программирования, обучения и применения на стандартном оборудовании. Предложенные алгоритмы применения СНС универсальны относительно архитектур со сверточными и подвыборочными слоями и могут использоваться для поиска нетекстовых объектов. Актуальными задачами являются совершенствование предложенной модели детектора с целью снижения количества ложных срабатываний (принятий фонового фрагмента за текстовый) и дальнейшая вычислительная оптимизация модели на базе векторизации.

СПИСОК ЦИТИРОВАННЫХ ИСТОЧНИКОВ

1. Кузьмицкий, Н.Н. Построение целостных контуров объектов на полутоновых изображениях / Н.Н. Кузьмицкий, С.С. Дереченник // Информационные технологии и системы: материалы Международной научной конференции. – 2011. – С. 175–176.
2. Krizhevsky, A. Imagenet classification with deep convolutional neural networks / A. Krizhevsky, I. Sutskever, G. Hinton // Proceedings of the International Conference on Neural Information Processing Systems. – 2012. –. Vol. 1. – P. 1097–1105.
3. LeCun, Y. Gradient-Based Learning Applied to Document Recognition / Y. LeCun, L. Bottou, Y. Bengio, P. Haffner // Proceedings of the IEEE. – 1998. – Vol. 86. – P. 2278–2324.
4. Кузьмицкий, Н.Н. Построение универсальных классификаторов текстовых образов русского языка на базе сверточных нейросетей / Н.Н. Кузьмицкий // Доклады БГУИР. – 2015. – № 4. – С. 33–39.
5. Калиновский, И.А. Обзор и тестирование детекторов фронтальных лиц / И.А. Калиновский, В.Г. Спицын // Компьютерная оптика. – 2016. – Т. 40, № 1. – С. 99–111.
6. Druki, A.A. Application of Convolutional Neural Networks for Automatic Number Plate Recognition on Complex Background Images / A.A. Druki, J.A. Bolotova, V.G. Spitsyn // Applied Mechanics and Materials. – 2015. – Vol. 756. – P. 695–703.
7. Wang, K. End-to-end scene text recognition / K. Wang, B. Babenko, S. Belongie // Proceedings of the IEEE International Conference on Computer Vision. – 2011. – Vol. 6. – P. 1457–1464.
8. Delakis, M. Text detection with convolutional neural networks / M. Delakis, C. Garcia // Proceedings of the International Conference on Computer Vision Theory and Applications. – 2008. – Vol. 2. – P. 290–294.
9. Sermanet, P. OverFeat: Integrated recognition, localization and detection using convolutional networks / P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun [Электронный ресурс]. – Режим доступа : http://arxiv.org/abs/1312.6229.pdf. – Дата доступа: 01.08.2017.
10. Garcia, C. Convolutional face finder: A neural architecture for fast and robust face detection / C. Garcia, M. Delakis // Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2004. – Vol. 26. – P. 1408–1423.
11. NeuroPro нейронные сети, методы обработки и анализа данных: от исследований до разработок и внедрений [Электронный ресурс]. – Режим доступа : http://neuropro.ru/memo312.shtml. – Дата доступа : 06.08.2017.
12. Open Source Computer Vision Library [Электронный ресурс]. – Режим доступа : https://github.com/opencv/opencv/blob/master/data/haarcascades/haarcascade_russian_plate_number.xml. – Дата доступа : 07.08.2017.
13. ICDAR 2013 Robust Reading Competition [Электронный ресурс]. – Режим доступа : http://rrc.cvc.uab.es/?ch=2&com=evaluation. – Дата доступа : 08.08.2017.

Материал поступил в редакцию 24.10.2017

KUZMITSKY N.N. Detection of text objects based on the «not-deep» convolutional neural network with optimization of calculations

Paper presents the model of text detector in form of «not-deep» convolutional neural network and the method of its application based on modified multiscale fragmentation of image, which reduces resource intensity of processing by more than two orders in comparison with standard fragmentation. The algorithm for text localization based on responses of detector is developed, which adaptability exceeds similar ones due to joint analysis of responses in adjacent lines and close scales of image, which allows localizing distorted text blocks of different sizes and orientations.

Based on the neural network model, the module for text detection is created, applicable for processing images with an arbitrary composition. Taking into account the priori information and features of the chosen software platform, ways of reducing resource intensity of the module are determined. Testing the module on sample of images reflecting moment of vehicles entry to protected area demonstrated high quality of registration numbers text localization, which exceeds level of the specialized module based on Haar cascade.

УДК 004.81

Крапивин Ю.Б.

ЛИНГВИСТИЧЕСКИЙ АНАЛИЗ ТЕКСТА В ЗАДАЧЕ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ ЗАИМСТВОВАННЫХ ФРАГМЕНТОВ ТЕКСТОВЫХ ДОКУМЕНТОВ

Введение. Постоянно увеличивающийся объем информации, представленной на различных языках как в полнотекстовых базах данных, так и в сети Интернет, обостряет проблему ее оперативной и качественной обработки с целью удовлетворения информационной потребности пользователей. Под информационным поиском (ИП) обычно понимают непосредственно процесс поиска и пред-

Крапивин Юрий Борисович, *старший преподаватель кафедры интеллектуальных информационных технологий Брестского государственного технического университета.*
Беларусь, БрГТУ, 224017, г. Брест, ул. Московская, 267.

ставления пользователю информации в соответствии с запросом, который, в свою очередь, и отражает эту информационную потребность. Одним из основных типов ИП является поиск релевантных документов, т. е. тех, которые схожи по содержанию с заданным документом-образцом, который и выступает в качестве запроса пользователя. Существенным достоинством такого запроса является его большая информативность, что способствует эффективному решению задачи. Оно обычно основывается [1] на некоторой процедуре индексирования, в результате которого с использованием, как правило, определенной лингвистической обработки строится формальное представление (поисковый образ) и запроса (ПОЗ), и документов (ПОД) из поисковой базы, а также процедуры сравнения поисковых образов запроса и документов с определением согласно некоторому правилу степени их соответствия (релевантности). На основе получаемых оценок принимается решение о выдаче или невыдаче того или иного документа.

Самое непосредственное отношение к задаче поиска документов, релевантных данному, имеет актуальная задача автоматического распознавания заимствования текстовых фрагментов (плагиата), который может быть явным (лексическим), когда речь идет об одном и том же фрагменте текста, принадлежащем разным текстовым документам (могут допускаться минимальные расхождения, например, за счет использования вводных слов, синонимов и т. п.), и неявным (семантическим), когда речь идет о фрагментах различных текстовых документов, имеющих одинаковый, по отношению к заданной системе знаний, но выраженный разными цепочками символов, смысл.

Проведенный анализ [2] показал, что существующие решения задачи автоматического распознавания заимствованных фрагментов (ЗФ) фактически ориентированы на распознавание лексически заимствованных фрагментов с учетом, в лучшем случае, простейших морфологических преобразований и отношений синонимии, не используют развитого лингвистического анализа текстовых документов и, следовательно, не ориентированы на решение задачи с учетом более сложных преобразований текста и автоматического распознавания семантически заимствованных фрагментов, тем более в многоязычной информационной среде.

С учётом вышеизложенного нами была разработана следующая принципиальная структурно-функциональная схема системы автоматического распознавания ЗФ (рис. 1).

Предполагается, что **поисковое пространство (ПП) включает в себя текстовые Интернет-документы и документы из поисковой БД пользователя.** В общем случае эти документы представлены на различных естественных языках (ЕЯ), т.е. ПП является многоязычной информационной средой. Входной информацией для системы являются подлежащие анализу на предмет заимствования Входной документ (блок 1), представленные в одном из наиболее распространенных форматов: TXT, RTF, DOC, DOCX, PDF, HTML. Ориентируясь на огромные объемы ПП и учитывая жесткие ограничения, накладываемые на время реакции системы автоматического распознавания ЗФ, этот документ сначала поступает в Подсистему поиска релевантных документов (блок 2). Ее задача – за минимальное время найти в используемом ПП подмножество наиболее релевантных входному текстовых документов, представленных, в общем случае, как на языке входного документа, так и на других ЕЯ, которые далее поступают в Подсистему собственно распознавания заимствованных фрагментов (блок 3) и затем – в Подсистему формирования отчета (блок 4) с последующим предоставлением его пользователю. Взаимодействие с системой осуществляется посредством Интерфейса пользователя (блок 5), который поддерживает ввод документов и просмотр результатов поиска заимствований.

Очевидно, что функционирование системы в режиме cross-language требует функциональности Подсистем определения языка текстового документа (блок 6) и машинного перевода (блок 7). При этом, функциональность последней из них, в зависимости от количе-

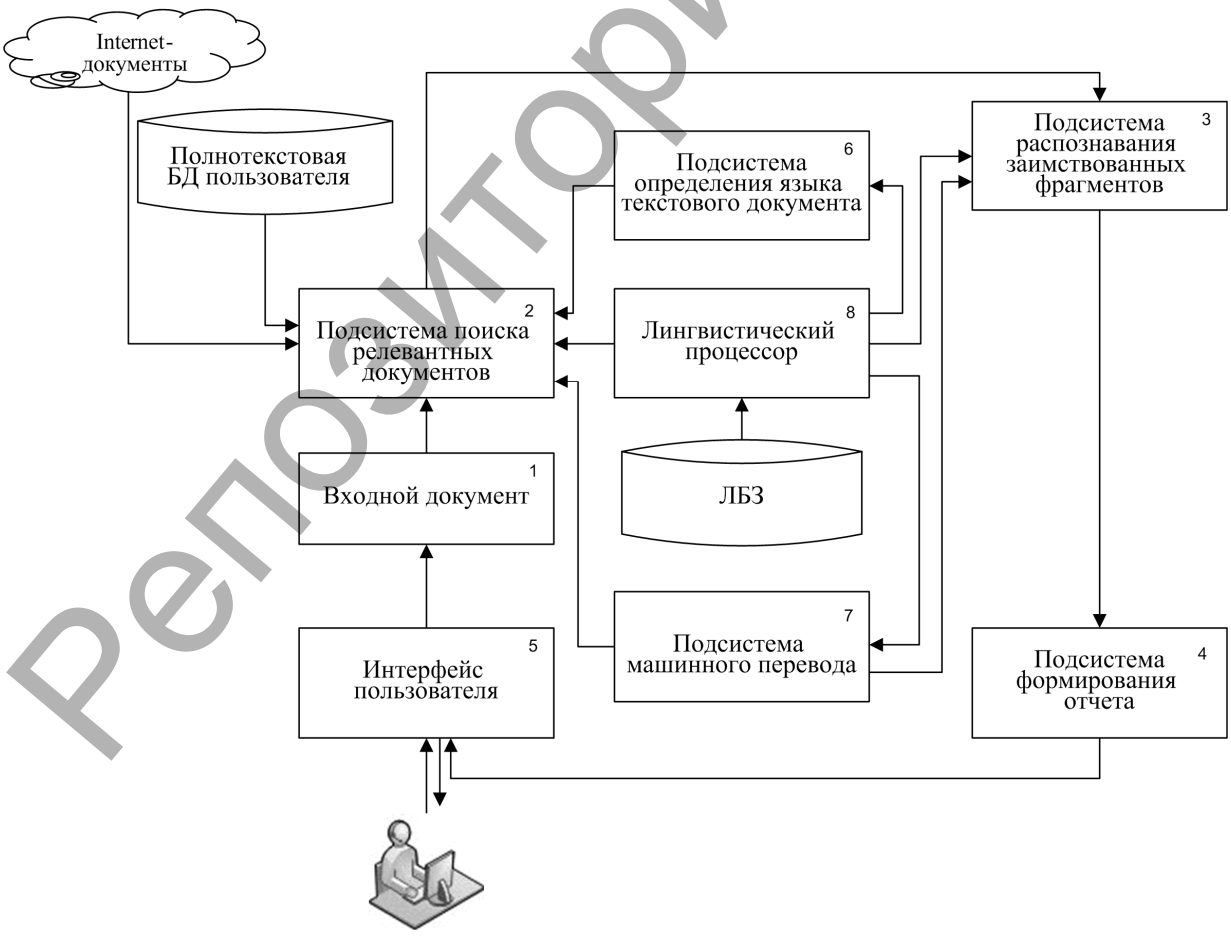

Рисунок 1 – Структурно-функциональная схема системы автоматического распознавания ЗФ

ства используемых языков и параметров распределения на них текстовых документов в ПП, ориентирована как на перевод текстов самих документов, так и их поисковых образов. Причем первый случай может иметь место как при поиске релевантных документов, так и на этапе распознавания ЗФ. Например, найденные на основе перевода ПОД релевантные Интернет-документы будут переводиться в блоке 3, а для документов из полнотекстовой БД пользователя с целью минимизации общего времени решения задачи в ряде случаев целесообразно заранее получить эквиваленты на других ЕЯ и хранить их в этой же БД. Отметим, что в том частном случае, когда имеет место одноязычная информационная среда, подсистема поиска релевантных документов не использует функциональность подсистем блоков 6 и 7. Более того, если ПП ограничивается только документами полнотекстовой БД пользователя и она является относительно постоянной и небольшой, от нескольких единиц до нескольких сотен документов, то эта подсистема «не включаясь» сразу передаст входной документ в Подсистему распознавания ЗФ. Что касается Лингвистического процессора (блок 8) и используемой им Лингвистической базы знаний, то они обеспечивают автоматический лингвистический, в том числе и лингвостатистический, анализ текстовых документов в той мере, в какой это необходимо для подсистем блоков 2, 3, 6, 7, что в свою очередь зависит от используемых там методов решения соответствующих задач. Можно говорить, что в задаче автоматического лингвистического анализа текстовых документов (блок 8), а также распознавания их языка (блок 6) и их машинного перевода (блок 7) составляют базовую лингвистическую обработку текстовых документов при решении задачи распознавания ЗФ.

Таким образом, в соответствии с приведённой схемой задача автоматического распознавания ЗФ, в целом, включает в совокупности следующие задачи: 1) поиска релевантных входному текстовых документов в полнотекстовой БД и в сети Интернет; 2) автоматического распознавания языка текстового документа; 3) машинного перевода текстовых документов и их поисковых образов во множестве заданных ЕЯ; 4) автоматического распознавания лексически и семантически заимствованных фрагментов текстовых документов; 5) автоматического построения отчета 6) автоматического лингвистического анализа текстовых документов.

С учетом приведенного перечня основных, подлежащих решению, задач в рамках общей задачи автоматического распознавания ЗФ, их анализа, а также основываясь на обоснованном тезисе об усилении лингвистической составляющей при построении решений этих задач, могут быть сформулированы требования к необходимому уровню автоматического лингвистического анализа текста (задача 6). Безусловно, что в этом плане наиболее серьезные требования выдвигает задача автоматического распознавания семантически ЗФ, поскольку здесь речь идет об уровне семантического анализа текста. А для этого, очевидно, необходима вся функциональность так называемого базового лингвистического процессора (ЛП) [3]:

- форматирования текста; на этом этапе проводится преобразование текстовых документов, представленных обычно в различных форматах, в некоторый единый формат, максимально сохраняющий стилистическую и структурную разметку документов; здесь же осуществляется разбиение текста на параграфы, выделение заголовков, подзаголовков и т. п.;
- лексического анализа текста, предназначенного, прежде всего, для распознавания в обрабатываемом тексте границ слов и предложений;
- лексико-грамматического анализа текста, его задачей является определение лексико-грамматического класса каждого слова входного текста с учетом контекста и заранее заданного списка этих классов для ЕЯ; так для английского языка их количество может достигать 200 (это, например [4], JJ – прилагательное; VB – глагол; MD – модальный глагол; NN – существительное единственного числа; RB – наречие; ATI – определенный артикль; CC – союз), а для белорусского и русского – более 1000;
- синтаксического анализа текста, на данном этапе традиционно для каждого предложения текста строится синтаксическое дерево, в котором его слова представлены в виде листьев, а другим узлам соответствует например простая именная, предложная, глагольная и т. п. группы, в том числе и собственно предложение (корень дерева). При этом устанавливаются синтаксические связи между ними. Например, фиксируется факт, что простое предложение включает подлежащее, выраженное именной группой, сказуемое, выраженное глаголом в прошедшем времени, и прямое дополнение, также выраженное именной группой.
- семантико-синтаксического анализа текста, его задачей является извлечение из синтаксических деревьев в виде отношений наиболее значимых, с точки зрения знаний основных типов [5], синтаксических структур: Simple Noun Phrase (простая именная группа), Verb Phrase (глагольная группа), Noun Phrase Additional (именная группа распространенная различными оборотами), Complex Sentence (сложноподчиненное предложение).

Основными компонентами лингвистической базы знаний (ЛБЗ) базового ЛП являются следующие:

- классификаторы лексико-грамматических, синтаксических и семантических свойств ЕЯ; их состав зависит от конкретных свойств ЕЯ и от характера приложения, определяющего степень детализации лингвистического анализа текста;
- базовый (эталонный) словарь, он реализуется в виде словаря словоформ ЕЯ и включает максимально возможное их количество, при этом для каждой словоформы указаны все ее возможные вне контекста лексико-грамматические классы (ЛГК);
- базовый (эталонный) корпус текстов (БКТ); реализуется в виде определенным образом подобранных текстов, причем, как минимум каждому слову текста указан его единственный с точки зрения контекста ЛГК; минимальный размер БКТ обычно составляет порядка одного миллиона словоупотреблений, однако моделирование ЕЯ на более высоких уровнях его глубины требует разработки БКТ объемом порядка $10^7 - 10^8$ словоупотреблений, причем аннотированных не только ЛГК, но и, возможно, метками синтаксических и семантических отношений; БКТ предназначен, прежде всего, для получения количественных оценок языка, тестирования лингвистических гипотез и отдельных алгоритмов и систем автоматической обработки текста;
- лингвистические правила анализа (ЛПР) текста на различных уровнях глубины ЕЯ; такие правила, получаемые лингвистами-экспертами, являются основой разработки машинных алгоритмов для большинства этапов автоматического лингвистического анализа текста; совокупность этих правил, например, для лексико-грамматического и синтаксического анализа составляет грамматику ЕЯ; их количество, в зависимости от глубины лингвистического анализа текста и степени обобщения терминальных символов, может колебаться от нескольких десятков до десятков тысяч.

С целью машинной обработки ЛП должна быть разработана некоторая нотация для формального описания этих правил, в которой они обычно и представляются в ЛБЗ. Причем, предлагаемый формализм должен быть максимально соотнесен с требованиями его доступности для использования экспертами, возможностью обобщения разрабатываемых правил и оптимизации скорости их обработки. В [3] в качестве такого формализма разработан и успешно внедрен в промышленные приложения так называемый язык расширенных регулярных выражений (WRE). В дополнение к основным перечисленным ресурсам в состав ЛБЗ входят различного рода словари (словари идиом, аббревиатур, имен собственных, слов, параметров и т. п.), специальные лексические базы данных, например, типа WordNet [6], отображающие синонимические иерархические и ассоциативные отношения концептов и т. д. Безусловно, ЛБЗ, ориентируясь на обработку текстов на нескольких языках, является многоязычной. В целом, функциональности представленного выше базового ЛП, очевидно, тем более достаточно для обеспечения лингвистической составляющей при решении задачи (1) – для автоматического распознавания ключевых слов с целью построения для текстовых документов их ПОДов, задачи (2) – для автоматического построения поискового образа языка, например, в виде словаря грамматических слов [7] и т. д., и задачи (4). Этот же базовый ЛП обеспечивает и требуемый для решения всех задач уровень лингвостатистического анализа текста, который традиционно сводится,

Таблица 1 – Фрагмент многоязычной лексической БД MModWN

Номер синсета	Определение	Синонимичные ряды		
		Английский	Французский	Немецкий
109459609	any mechanical force that tends to retard or oppose motion	• RESISTANCE	• RÉSISTANCE	• WIDERSTAND
109421558	the resistance encountered when one body is moved in contact with another	• FRICTION • RUBBING	• FRICTION • FROTTEMENT	• REIBUNG • FRIKTION
109421888	the process of wearing down or rubbing away by means of friction	• ABRASION • ATTRITION • GRINDING	• ABRASION • BROYAGE • MEULAGE	• ABRIEB • ABNUTZUNG • VERSCHLEI • SCHLEIFEN • SCHLIFF

прежде всего, к подсчету частот, распознаваемых на приведенных выше этапах обработки текста лексических единиц и отношений. Остановимся подробнее на задаче 3) – машинного перевода текстовых документов и их поисковых образов.

Согласно структурно-функциональной схеме системы автоматического распознавания ЗФ, приведенной на рисунке 1, ориентированной в общем случае на огромные объемы ПП, как относительно интернет-документов, так и документов из полнотекстовой БД пользователя, необходимо каким-то образом предварительно быстро и эффективно, с точки зрения релевантности входящих в ПП документов данному, минимизировать это пространство, после чего управление предается подсистеме распознавания ЗФ (блок 3), которая и решает целевую задачу путем «сплошного» сравнения входного документа с документами из «минимизированного» ПП (МПП). Т.е. именно процедура поиска релевантных документов является механизмом минимизации ПП. Как уже отмечалось ранее, речь при этом идет о многоязычной информационной среде и о cross-language функциональности. Первое предполагает, что если на входе задан текстовый документ на языке $L_i \in L$, где L – множество ЕЯ, с которыми поддерживает работу система, то МПП будет содержать множество всех документов из ПП, релевантных данному и представленных на этом же языке L_i. Функциональность машинного перевода в данном случае, очевидно, не нужна. Второе предполагает, что при том же заданном документе МПП будет содержать все, независимо от языка представления, релевантные документы из ПП. И здесь уже без указанной функциональности, безусловно, не обойтись. Она в любом случае необходима для МП входного текстового документа на каждый ЕЯ из $L' \subseteq \overline{L}$ и, возможно, для МП его поискового образа на каждый ЕЯ из $\overline{L}$, где $\overline{L}$ – множество всех ЕЯ из L, кроме языка представления входного текстового документа или его поискового образа. Причем, использование последнего делает весь алгоритм решения задачи значительно менее трудоемким. Что касается процедуры поиска релевантных документов, то для полнотекстовой базы пользователя, которая ему вполне «прозрачна», целесообразно использовать уже существующие эффективные решения задачи такого поиска [8], а для случая Internet-документов к тому же должно быть предложено решение, жестко ориентированное на особенности механизмов, реализованных в современных поисковых машинах (одно из таких решений представлено нами в [9]). Как оказалось, независимо от типа ПП эти решения основаны на автоматическом распознавании в текстовых документах ключевых слов с целью построения их поисковых образов. И, таким образом, если для МП текстовых документов могут использоваться, как отмечалось ранее, уже существующие практические решения, то для МП их поисковых образов решение должно быть предложено.

Как показали проведенные исследования, основу МП поискового образа входного документа (запроса), ПОЗа, представленного списком ключевых слов, могут составить определенные двуязычные словари (словари концептов, акций и их атрибутов). Более того, такой подход позволяет эффективно решить задачу автоматического распознавания ЗФ даже на семантическом уровне. Указанные компоненты словарей, с одной стороны, являются уникальными семантическими понятиями, которые, в принципе, от языка не зависят и, таким образом, выступают в роли интерлингвы, а с другой – они являются «носителями» ключевых слов. Для хранения таких словарей может быть использована очень эффективная структура, а именно структура многоязычной лексической БД MModWN, аналогичная по своей структуре WordNet [6], которая описывает концепты внешнего мира в форме пронумерованных понятий (синсетов), выраженных набором синонимичных слов и словосочетаний на всех языках из множества L, а также различные семантические отношения между концептами («общее-частное», «часть-целое», «группа-элемент» и т. д.). При этом сохраняются, очевидно, существовавшие бинарные отношения между лексическими единицами и достигается их минимальная избыточность.

В работе [10] описана общая технология построения многоязычной лексической БД MModWN и технология построения двуязычных словарей, основанная на автоматической обработке параллельных корпусов текстов и распознавании в них так называемых параллельных фактов, а также их компонентов и атрибутов.

Идея использования структуры описанной выше многоязычной лексической БД для трансляции ПОЗа входного текстового документа на языке L_i в множество ПОЗов на языках из $\overline{L}$ значительно упрощает эту процедуру, сводя её к поиску для каждого ключевого слова L_i – ПОЗа его синсета и последующему выбору из него представленных там его синонимов, а также эквивалентов на языках из $\overline{L}$. Так, ниже приводится фрагмент указанной лексической БД MModWN на примере трёх ЕЯ.

Причем, существующая единая для всех языков нумерация понятий позволяет сохранить все установленные семантические отношения между ними. Так, например, семантические отношения «общее-частное» между пронумерованными понятиями 109459609→109421558 и 109421558→109421888 определяют отношения между соответствующими им синонимичными рядами в различных языках. Тогда, если, например, L_i – ПОЗ на английском языке включает наряду с другими ключевое слово ABRASION, то, в соответствии с приведенным выше фрагментом БД MModWN, в L_i – ПОЗе оно будет дополнено синонимами ATTRITION и GRINDING, а при переводе на французский язык получит три эквивалента (ABRASION, BROYAGE, MEULAGE) и пять эквивалентов – при переводе на немецкий язык (ABRIEB, ABNUTZUNG, VERSCHLEI, SCHLEIFEN, SCHLIFF).

Что касается последующей работы системы автоматического распознавания ЗФ в рассматриваемом режиме cross-language с уже МПП, то требуемая здесь функциональность МП, как показали проведённые исследования и опыт разработки конкретной системы [11], может быть обеспечено, например, существующими качественными системами МП (в нашем случае это была белорусско-русская информационная среда и соответственно система [12]). В противном случае можно перейти к решению задачи на семантическом уровне, рассматривая сравниваемые фрагменты не как цепочки слов, а как

цепочки фактов, представленных концептами, акциями и их атрибутами и распознаваемых ЛП с использованием лексической БД, аналогичной упомянутой выше БД MModWN [13].

Заключение. Привлечение средств развитого лингвистического анализа текста, опирающихся на знания о ЕЯ, реализация cross-language функциональности за счёт механизма дополнения ключевых слов ПОЗа с помощью многоязычной лексической БД, обеспечивающей транслитерацию составляющих ПОЗа на основании учёта семантических отношений между их переводными эквивалентами, позволяет быстро и эффективно минимизировать ПП, тем самым не только обеспечивая возможность качественного решения задачи автоматического распознавания ЗФ на лексико-грамматическом, но и на семантическом уровне ЕЯ.

СПИСОК ЦИТИРОВАННЫХ ИСТОЧНИКОВ

1. Manning, C. Introduction to Information Retrieval / C. Manning, P. Raghavan, H. Schütze. – 1 edition. – Cambridge University Press, 2008. – 496 p.
2. Крапивин, Ю.Б. К задаче автоматического распознавания воспроизведенных фрагментов текстовых документов / Ю.Б. Крапивин // Вестник БрГТУ : Физика, математика, информатика. – 2009. – № 5(59). – С. 120–123.
3. Чеусов, А.В. Разработка алгоритмов и технологии построения многоязычного базового лингвистического процессора : дис. канд. тех. наук. / Чеусов А.В. – Минск, 2013. – 116 с.
4. Lancaster-Oslo-Bergen Corpus [Электронный ресурс]. – 2017. – Режим доступа: http://www.hit.uib.no/icame/lobman/lob3.html / – Дата доступа : 17.09.2017.
5. Совпель, И.В. Автоматическое распознавание основных типов знаний в текстовых документах / И.В. Совпель // Искусственный интеллект. – 2007. – № 3. – С. 328–332.
6. WordNet [Электронный ресурс]. – 2017. – Режим доступа : http://wordnet.princeton.edu/ – Дата доступа : 17.09.2017.
7. Крапивин, Ю.Б. Автоматическое определение языка текстового документа для основных европейских языков / Ю.Б. Крапивин // Информатика. – 2011. – № 31 июль-сентябрь. – С. 112–116.
8. Мамчич, А.А. Модели и алгоритмы информационного поиска в многоязычной среде на основе тематических и динамических корпусов текстов: дис. канд. тех. наук. / А.А. Мамчич – Минск, 2011. – 122 с.
9. Крапивин, Ю.Б. Автоматический поиск заимствованных из Интернет-источников фрагментов / Ю.Б. Крапивин // Искусственный интеллект. – 2012. – № 4. – С. 183–189.
10. Постаногов, Д.Ю. К вопросу многоязычности систем инженерии знаний и их приложений / Д.Ю. Постаногов, И.В. Совпель // Искусственный интеллект. – 2006. – № 3. – С. 474–479.
11. Отчет о научно-исследовательской работе «Разработать технологию и инструментально-программный комплекс распознавания в диссертационных работах случаев заимствования без ссылок на авторов» – Минск, 2009. – 51 с.
12. Воронков, Н.В. Методы, алгоритмы и модели систем автоматического реферирования текстовых документов: дис. канд. тех. наук / Н.В. Воронков – Минск, 2007. – 165 с.
13. Крапивин, Ю.Б. Функциональность cross-language в задаче автоматического распознавания семантически эквивалентных фрагментов текстовых документов / Ю.Б. Крапивин // Искусственный интеллект. – 2013. – № 4. – С. 187–194.

Материал поступил в редакцию 26.12.2017

KRAPIVIN Yu.B. The linguistic analysis of the text in a problem of automatic recognition of the borrowed fragments of text documents

The requirements for the needed level of the automatic linguistic analysis of the text with the purpose of automatic recognition of the adopted fragments of the text documents were formulated. Methods of qualitative improvement of the solving the problem via the usage of facilities of specialized linguistic resources are proposed. They afford to perform the analysis up to semantic-syntactic level of the language.

УДК 338.2:681.3

Матюшков А.Л, Матюшкова Г.Л.

НЕЙРОННАЯ СЕТЬ ДЛЯ УСТАНОВЛЕНИЯ РЕЙТИНГА ОБЪЕКТА

Введение. В прикладном плане знание рейтинга объекта широко используется при покупке различного оборудования, программных продуктов, рекламе и т. д.

Автоматизация установления рейтинга объекта позволяет обеспечить недостающим инструментом задачу комплексного принятия решения.

Современные методы принятия решений широко начинают использовать различные типы сетей [1, 2, 3], включая их настройку и обучение на специфику решаемых задач.

Многие объекты иногда необходимо расположить в порядке возрастания их рейтинга или выбрать из них более подходящий в порядке его убывания.

Основой для решения задачи часто служат численные оценки, расположенные в произвольном порядке.

Таким образом, после выполнения специфических расчётов задача сводится к присвоению каждому члену ряда (результату) своего номера (рейтинга) относительно объявленного критерия важности (простейший случай – цена, вес, объём, площадь и т. п.). Далее в порядке рейтинга выбирают приемлемый для заказчика объект.

Эту задачу можно свести к присвоению членам положительного ряда номеров с целью установления их значимости с помощью многослойной нейронной сети всего из шести нейронов при её обучении методом обратного распространения ошибок, описанного с рекомендациями по выбору параметров сети в [1]. Нами добавлена следующая модификация: указана конкретная активационная функция $y_i=1/(1+e^{-x_i})$ для всех нейронов скрытых слоёв и предложена своеобразная функция смещения для определения результата на выходе сети как ближайшего целого числа к положительному аргументу, что позволило получать рейтинг в требуемой форме, обучение сети завершается при нулевой ошибке.

Чтобы показать характер особенностей применения нейросетевых методов, проиллюстрируем их использование для поиска рейтинга любого из 6 заданных объектов по их описаниям в виде ряда положительных действительных чисел с помощью многослойной нейронной сети (рис. 1).

Её структура определяется из характера задачи:

- первый слой включает нейроны 1 и 2 входы;
- второй (скрытый) слой включает нейроны 3, 4, 5;
- и третий слой включает нейрон 6 выход.

Матюшков Александр Леонидович, *к.т.н., доцент Белорусского государственного университета информатики и радиоэлектроники. Беларусь, БГУиР, 220013, г. Минск, ул. П. Бровки, 6.*

Матюшкова Галина Леонидовна, *научный сотрудник ОИПИ НАН Беларуси. Беларусь, 220012, г. Минск, ул. Сурганова, 6.*