УДК 81’322.+81’33+811.58
ББК Ш.11+Ш171

ПРИКЛАДНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ТЕКСТОВ НА ВОСТОЧНЫХ ЯЗЫКАХ: СОВРЕМЕННЫЕ ПЕРСПЕКТИВЫ

Б.Г. Фаткулин

Статья посвящена развитию прикладной лингвистики в области востоковедения, обработке естественных языков при помощи компьютерных алгоритмов и программ. В целях развития поликультурного языкового образования предлагается расширение списка преподаваемых в России языков. Раскрываются основные типы востоковедческих исследований в России. Анализируются возможности современного программного обеспечения в области обработки восточных языков. В качестве примера показан опыт по извлечению терминологии из текстов на китайском языке с помощью инструментов прикладной лингвистики.

Ключевые слова: обработка естественного языка, анализ текста, извлечение информации, извлечение определений терминов, обработка естественного языка, выделение ключевых словосочетаний, интеллектуальный анализ текстовых данных, востоковедение, восточные языки.

В последнее время активно обсуждаются проблемы развития инновационных образовательных моделей, отвечающих критериям многополярного мира. Стремление преодолеть однополярность вызвало появление таких международных и региональных объединений как Шанхайская организация сотрудничества (ШОС), Азиатско-Тихоокеанское экономическое сотрудничество (АТЭС), группа стран БРИКС, а также сетевых университетов ШОС, Институтов Конфуция [3, с. 255].

Многополярность мира должна сопровождаться и многовекторностью информации. Многополярный подход к информации требует формирования профессионального сообщества и системы профессиональных коммуникаций, подготовки кадров специалистов-аналитиков. Соответственно, возрастает роль востоковедения (ориенталистики) – совокупности научных дисциплин, изучающих историю, экономику, языки, литературу, этнографию, искусство, религию, философию стран Востока, включающую региональные отрасли (напр., арабистика, синология, иранистика, кореевед ение, японоведение, афганистика, тюркология, индология и т. д.).

Исходя из предъявляемых к специалисту требований, преподавание иностранных языков в системе высшего образования должно исключить однозначное преобладание западноевропейских языков. В основу подготовки специалистов должна быть положена поликультурная модель высшего образования, которая зиждется на тезисе о том, что европейские языки дают только один вариант из множества возможных вариантов картины мира.

Востоковеды-аналитики ведут работу с источниками информации на языках регионов, находящихся в фокусе их внимания. Для переработки больших объемов информации специалисты должны владеть методами прикладной лингвистики.

Задачи, стоящие перед современным востоковедением, определяются тем, что оно делится на два больших направления:

• архивное востоковедение, занимающееся историческими свидетельствами древних восточных цивилизаций, архивными материалами на восточных языках: рукописями, манускриптами.

• аналитическое востоковедение, рассматривающее восточные страны как субъекты международной политики, считающее восточные народы и страны неотъемлемой частью современной человеческой цивилизации. Аналитическое востоковедение использует современные тексты на восточных языках с целью сбора информации, написания аналитических статей, справок, прогнозов.

Несмотря на некоторую разницу в предмете и объекте, обе ветви востоковедения вплотную занимаются текстами на восточных языках, и поэтому для обеих ветвей актуальна прикладная лингвистика, дающая ключ к эффективной и быстрой обработке больших массивов данных [6, с. 111], та ее часть, которая в международной науке называется Natural Language Processing.

Необходимо развивать профессиональную подготовку лингвистов-прикладников в области восточных языков, которая бы подчинялась следующим задачам:

1) программно-аналитическая (нахождение закономерностей [5, с. 160] в восточных языках, которые помогают разрабатывать программное обеспечение, разработка прикладных утилит для обработки текстов на восточных языках, локализация программного обеспечения [2:25], направленного для реализации на восточных рынках);

2) организационно-идеологическая (пропаганда в странах Востока русского языка, российских традиционных ценностей, российских геополитических интересов);

3) лексиколого-терминографическая (обновление и пополнение тезауруса востоковедческих дисциплин современной терминологией на восточных языках, лексикографический мониторинг терминологии, появляющейся в разных отраслях науки и культуры в государствах Востока, использование достижений прикладной лингвистики в работе журналистов и публицистов, пишущих на «восточные темы»);

4) учебно-методическая (разработка российского программного обеспечения в области преподавания восточных языков, создание пособий).

Объектами прикладной лингвистики в области востоковедения являются следующие восточные языки:

– Дальневосточные языки:

• китайский (один из самых актуальных в силу возрастания объемов российско-китайского сотрудничества во всех сферах)

• корейский (в виду того, что Южная Корея находится «на гребне волны» технологического прогресса и имеет высокий потенциал для научно-технического сотрудничества)

• японский, хотя его позиции в России осложняются тем, что страна является союзником США и находится в орбите американского культурного влияния.

– Средневосточные языки (языки современных Ирана и Афганистана): фарси, дари, пушту. Эти языки являются системообразующими языками регионов мира, непосредственно входящих в сферу жизненных интересов РФ.

– Ближневосточные языки (арабский язык). Позиции арабского языка ослаблены тем, что в настоящее время арабские страны переживают сложный период своего развития.

– Турецкий язык, сфера применения которого ограничена тем, что Турция является страной НАТО и соперничает с Россией за влияние в Закавказье.

Кроме того, возрастает потребность в разработке научных подходов к обработке текстов на государственных языках республик, в свое время входивших в состав СССР.

По всем вышеперечисленным языкам необходимо готовить кадры в области прикладной лингвистики и обработке естественных языков [1, с. 15].

Первые проекты по обработке текстов на восточных языках появились еще в 90-х годах XX века, и сейчас эта отрасль порождает множество научных школ и развивается высокими темпами. Проекты Natural Language Processing разрабатываются как в США и в странах ЕС [8, с. 7; 9, с. 100], так и в КНР [7, с. 30].

Парадигма исследований по восточным языкам носит стандартный характер:

• создание и разметка корпусов на восточных языках (NLP Resources);

• создание прикладных утилит: сегментеров, парсеров, программ распознавания речи, программ распознавания символов и иероглифов, инструментов Named Entity Recognition.

Прикладная лингвистика в области восточных языков не может обойтись и без гуманитарной востоковедческой составляющей. Без историков, философов и «классических» востоковедов, которые будут поставлять контент, программное обеспечение само по себе останется просто набором алгоритмов.

В заключение приведем пример использования достижений прикладной лингвистики в востоковедении. Изучая терминологию ШОС, мы обратили внимание на такую сторону информации в документации ШОС, как именованные сущности. Именованные сущности – персоны, организации, географические объекты и прочие объекты, обозначаемые в тексте с использованием имен собственных. Named Entity Recognition (извлечение сущностей) – это одна из задач text mining, суть которой состоит в автоматическом определении сущностей в неструктурированных тестовых данных. Классическими сущностями выступают имена людей и компаний (names), адреса (locations), географические объекты (locations), даты (dates) и, в более сложных случаях, связи между ними, а также события, причинно-следственные связи, хронометраж событий. Также можно добавить к этому списку такие сущности, как электронные адреса, телефоны, определенные типы данных (например, IP адреса). Таким образом named entities составляют одну из основ аналитики.

В современной прикладной лингвистике существуют многочисленные методы извлечения терминологии из больших массивов текстов, называемых корпусами.

Работа делилась на два этапа:

1) подготовка корпуса текстов по ШОС;

2) выделение имен собственных с помощью алгоритмов Named Entity Recognition.

На первом этапе работы мы решили собрать репрезентативный корпус статей ШОС на китайском языке. Для этого при помощи программы wget рекурсивно скачали китайский вариант сайта infoshos.ru. Затем мы выбрали из полученных файлов только те файлы, которые содержали текстовую информацию. После этого мы объединили полученные тексты в один большой файл и подвергли его обработке при помощи программы Stanford Named Entity recognition. Кроме того, мы проверили наличие выделенных именованных сущностей в программе Wikimeta.

Полученные одиночные термины или терминологические словосочетания выводились из текста и переводились при помощи китайско-русского электронного переводчика с последующим ручным редактированием и поиском эквивалентов.

В результате применения вышеназванных методик исследования нами был составлен список терминов, используемых в освещении деятельности ШОС на китайском языке, включающий 52 понятия.

Мы использовали достижения прикладной лингвистики для поиска и выделения «именованных сущностей» ШОС: названий структур ШОС, персоналий руководящих работников, мероприятий в рамках этой организации, основных формулировок и подходов, встречающихся в ее руководящих документах. Между тем, в документах ШОС содержится очень много оригинальных концептов, описывающих взаимоотношения России и республик Средней Азии, Китая и России.

Таким образом, возрастающая роль восточных языков в многополярном мире, возрастающие объемы информации на восточных языках требуют повышенного внимания со стороны прикладной лингвистики. На наш взгляд, эта научная тематика обладает несомненными перспективами.

Литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика / Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ и др. – М.: МИЭМ, 2011. – 272 с.

2. Власов, Д.Ю. Извлечение отношений между понятиями из текстов на естественном языке / Д.Ю. Власов, Д.Е. Пальчунов, П.А. Степанов // Вестн. Новосиб. гос. ун-та. Серия «Информационные технологии». – 2010. – Т. 8. – Вып. 3. – С. 23–33.

3. Гурулева, Т.Л. Высшее языковое образование в России: восточный вектор / Т.Л. Гурулева // Известия РГПУ им. А.И. Герцена. – 2008. – № 71. – С. 252–261.

4. Соколовский, А.Я. Подготовка специалистов по Южной и Юго-Восточной Азии в Восточном Институте ДВГУ / А.Я. Соколовский // Известия Восточного института. – 2007. – № 14.

5. Степанов, П.А. Системы анализа текстов естественного языка / П.А. Степанов // Альманах современной науки и образования. – 2013. – № 6 (73). – С. 159–161.

6. Степанов П.А. Автоматизация обработки текстов естественного языка // Вестник Новосиб. гос. ун-та. Серия «Информационные технологии». – 2013. – Т. 11, вып. 2. – С. 109–115.

7. Introduction to Chinese Natural Language Processing / Kam-Fai, Wong, Wenjie Li, Ruifeng Xu, and Zheng-sheng Zhang // Synthesis Lectures on Human Language Technologies. – 2009. – vol. 2, no. 1. – P. 1–148.

8. Zitouni, I. Natural language processing of semitic languages / I. Zitouni. – Berlin: Springer, 2014.

9. Jadidinejad, A.H. Evaluation of Perstem: A Simple and Efficient Stemming Algorithm for Persian / A.H. Jadidinejad, M. Fariborz and J. Dehdari. // In C. Peters, G.D. Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peñas, and G. Roda, editors, Multilingual Information Access Evaluation I. Text Retrieval Experiments, vol. 6241 of Lecture Notes in Computer Science. – Springer, Heidelberg, 2010. – P. 98–101.

Фаткулин Булат Гилимдарович, кандидат филологических наук, доцент кафедры общей лингвистики, Южно-Уральский государственный университет (Челябинск), bfatkulin@gmail.com

Поступила в редакцию 30 июня 2014 г.

Bulletin of the South Ural State University
Series "Linguistics"
2014, vol. 11, no. 3, pp. 15–18

APPLIED LINGUISTICS AND TEXT PROCESSING IN ORIENTAL LANGUAGES: MODERN PROSPECTS

B.G. Fatkulin, *South Ural State University, Chelyabinsk, Russian Federation, bfatkulin@gmail.com*

The problem of the identification of text parameters having negative impact on the quality of machine translation is considered in the article. Definition and classification of translatability indicators is given. Linguistic phenomena in the scientific texts in Russian are considered at the graphic, lexical and syntagmatic levels. The special attention is paid to textual features causing problems for machine translation at the syntagmatic level, including both universal features and ones specific to Russian. A classification of formal machine translatability indicators for Russian scientific texts is compiled on the basis of language features analysis. The distinguished classes are matched against text translation

problems associated with the corresponding indicator. The obtained results can be applied in translation practice and when developing tools for computational support of translation.

Keywords: machine translatability, translatability indicator, machine translation, scientific text, corpus.

References

1. Bol'shakova E.I., Klyshinskij Je.S., Landje D.V. *Avtomaticheskaja obrabotka tekstov na estestvennom jazyke i komp'juternaja lingvistika* [Automatic processing of natural language texts and computational linguistics]. Moscow, MIJeM Publ., 2011, 272 p.

2. Vlasov D.Ju., Pal'chunov D.E., Stepanov P.A. Izvlechenie otnoshenij mezhdu ponjatijami iz tekstov na estestvennom jazyke [Extracting relationships between concepts from natural language texts]. *Vestn. Novosib. gos. un-ta. Ser. Informacionnye tehnologii,* 2010, vol. 8, iss. 3, pp. 23–33.

3. Guruleva T.L. Vysshee jazykovoe obrazovanie v Rossii: vostochnyj vektor [The linguistic Higher Education in Russia: the eastern vector]. *Izvestija RGPU im. A.I. Gercena*. 2008, no. 71.

4. Sokolovskij A.Ja. Podgotovka specialistov po Juzhnoj i Jugo-Vostochnoj Azii v Vostochnom Institute DVGU [Training of Specialists in South and Southeast Asia in the Far East Studies Institute]. *Izvestija Vostochnogo instituta*. 2007, no. 14.

5. Stepanov P.A. Sistemy analiza tekstov estestvennogo jazyka [Systems of natural language texts analysis]. *Al'manah sovremennoj nauki i obrazovanija*. Tambov: Gramota, 2013, no. 6 (73), pp. 159–161.

6. Stepanov P. A. Avtomatizacija obrabotki tekstov estestvennogo jazyka [The Automatization of natural language texts processing]. *Vestn. Novosib. gos. un-ta. Serija: Informacionnye tehnologii.* 2013, vol. 11, vyp. 2. pp. 109–115.

7. Kam-Fai Wong, Wenjie Li, Ruifeng Xu, and Zheng-sheng Zhang Introduction to Chinese Natural Language Processing Synthesis Lectures on Human Language Technologies, 2009, vol. 2, no. 1, pp. 1–148.

8. Zitouni, Imed. *Natural language processing of semitic languages*. Berlin: Springer, 2014.

9. Jadidinejad Amir Hossein, Fariborz Mahmoudi, and Jon Dehdari. Evaluation of Perstem: A Simple and Efficient Stemming Algorithm for Persian. In Peters C., Nunzio G. D., Kurimo M., Mandl T., Mostefa D., Peñas A., and Roda G., editors, Multilingual Information Access Evaluation I. Text Retrieval Experiments, vol. 6241 of Lecture Notes in Computer Science. Springer, Heidelberg, 2010, pp. 98–101.

Bulat G. Fatkulin, candidate degree in philology, associate professor, Linguistics chair lecturer, South Ural State University (Chelyabinsk), bfatkulin@gmail.com

Received 30 June 2014