Terminology Extraction from the Baidu Encyclopedia

Bulat Fatkulin[1,2]

[1] South Ural State University, Chelyabinsk, Russia
[2] Chelyabinsk State University, Chelyabinsk, Russia
bfatkulin@gmail.com

Abstract. The article examines the use of the applied linguistics technologies in the teaching of orientalistics in the Russian Federation higher education system. The research discloses the methods of the terminological units extracting using texts in Chinese studies of the modern Afghanistan. The author shows the solutions for intensive summarization and annotation of Chinese texts and language teaching methods for students to work with assistive software. The achievements of the Stanford NLP group are used for the Chinese text segmentation and named entity recognition.

Keywords: terminology extraction, orientalistics, natural language processing.

The applied linguistics can not be a "thing-in-itself", it serves determined interests. The teaching of Oriental and Asian languages occupies the special place in the system of Russian higher education. As a rule, the orientalistics education is obtained by students in elder age, being combined with certain tasks. Elitism and interdisciplinarity are key features of Orientalistics.

Variety of oriental cultures surrounding Russia is reflected in a wide range of Orientalistic branches (Iranian studies, Arabic studies, Turkology, Indology, Afgan studies etc.) The Sinology occupies a leading position among them [1]. All major world civilization centers, including Russia and China, have their own versions of orientalistics branches and use their own terminology. The following reasons make Afghan studies in China actual:

1. In recent years China has been active in developing countries in Asia and Africa and is a main investor in Afghanistan.
2. Afghan Studies became of great significance because of the region strategic location, since China relies heavily on the oil resources of the Middle East. The strategy of the Silk Road revival require control of eastern transport corridors.
3. The knowledge of the Islamic world terminology (including Afghan culture) is necessary to struggle against the religious extremism and the Uighur separatism.

China has its own political doctrine, information sources and media, the Chinese Internet is governed by the political doctrine of the country and provides information in accordance with its interests. Typological structure of the Chinese language and hyeroglyphs makes the direct borrowing of political terminology of the European languages impossible, and this feature deprives the external actors their possibility for public opinion manipulation. The Soviet heritage Middle Asian republics, Afghanistan, Pakistan and Iran are the closest western neighbors of China [2]. Due to many internal and external factors (ongoing civil war, the presence of the foreign military forces, drug trafficking, etc.) Afghanistan is a subject of attention to both Russia and China.

Russia should be aware it's Chinese ally projects, and therefore the study of the peculiarities of the Chinese terminology of Afghan studies can enrich Russian analytical networks with valuable experience. Studying of the Chinese Afghanistics terminology is necessary [9] for professionals involved in the work of intergovernmental organizations such as the SCO, BRICS, Custom Union, etc.

It should be stressed that Chinese experts use in their Afghan studies their own authentic terminology which largely differes from the terminology of the English-speaking global network structures and it's equivalents in European and Russian languages [8]. The philosophy of Confucianism and theoretical and methodological approaches of the Communist Party of China form the base of Chinese political terminology and are unknown for the wide range of scientists who do not understand Chinese.

If we intend to collect the relevant information we have to handle in a short time a large number of texts in the original language. Bare translation of Orientalistics articles from Western magazines impoverishes informational awareness. Qualified orientalists should be able to work independently with Chinese Oriental sources and must apply innovative educational technologies [4]. Fast annotation and summarization of texts is required from students. The access to these technologies develops the creative potential of them. Russian Terminography should work at the intersection of Chinese Studies [6], Islamic Studies,Arabic Studies and Iranian Studies. The high complexity of the Chinese texts processing necessitates the use of innovative technologies, which are based on the latest achievements of applied linguistics.

There are numerous methods of terminology extraction from large amounts of text, called corpora. The variety of algorithms and programs in different programming languages are used to exfor the term extraction. There are software products also ready-to-use for ordinary researchers. We have found a lot of information about algorythms of applied linguistic programs in articles of a Tomsk famous explorer O. S. Jacko [10].

Applied Linguistics for Chinese includes a wide range of specialized programs such as:

1. segmenters,
2. morphological analyzers,
3. parsers,
4. converters of encodings,

5. characters OCR systems,
6. databanks [3].

The text segmentation is automatically produced by the segmenter — a special program or script. Character is determined by segmenter task to get some information from the text analysis parameters are set in advance. The joined information is provided in a certain manner and conducted in one of the programming languages. Three phases are logical segmenting process stages: first it is punctual collection of information, for example, it may be a code web pages. Then, it is data analysis, processing and transformation into the desired format. Finally — it is providing result output.

In our work we used tools such as:

1. Stanford Chinese segmenter `http://nlp.stanford.edu:8080/parser/`
2. Shanghai Chinese language segmenter `http://hlt030.cse.ust.hk/research/c-assert/`
3. Automatic annotation of Chinese texts `http://www.chinese-tools.com`

It is difficult to overestimate the advantages of parser using for fast processing of the Chinese text are. The segmenter makes grouping of characters into combinations. The essence of this phenomenon can be explained by comparing the presentation of texts in Russian and Chinese. In Russian, the words are separated by spaces. However terminological combinations usually consist of several words. The stable combinations of words are easily recognized by native Russian languagers, but such grouping of words is difficulty for foreigners. In Chinese texts similar gaps stay between the standard characters. But the Chinese word can consist of multiple characters [5]. Terms, in turn, may consist of several words. Segmenter solves the problem of putting a space between characters groups, allows you to find the terms of several groups of characters.

To carry out the above-mentioned routine operations related to the recovery terminology, we used the Stanford Chinese segmenter, which uses probabilistic algorithms. The program is designed by Pi-Chuan Chang, Huihsin Tseng and Galen Andrew. we downloaded and installed this segmenter on a personal computer running operating system Linux Ubuntu. It works in Java 6 (JDK1.6)

Two segmentation models are provided. The "ctb" model was trained with Chinese treebank (CTB) segmentation, and the "pku" model was trained with Beijing University's (PKU) segmentation. PKU models provide smaller vocabulary sizes and OOV rates on test data than CTB models.

For both CTB and PKU, we provide two models representing slightly different feature sets. Models "ctb" and "pku" incorporate lexicon features to increase consistency in segmentation. The details of the segmenter can be found in the paper [12]. The description of the lexicon features can be found in [13].

The program runs from the command line by means of this command:

```
segment.sh [-k] [ctb | pku] <filename> <encoding> <size>
 ctb: Chinese Treebank
 pku: Beijing Univ.
```

The main principle of the Stanford segmenter is described in the work of Levy and Manning [11].

The Chinese text before the processing looked as follows:

比尔兼德高地，北部有厄尔布兹山脉，德马万德峰海拔5670米，为伊朗最高峰。西部和 西南部是宽阔的扎格罗斯山山系，约占国土面积一半。中部为干燥的盆地，形成许多沙 漠，有卡维尔荒漠与卢特荒漠，平均海拔1，000余米。仅西南部波斯湾沿岸与北部里海 沿岸有小面积的冲击平原。西南部扎格罗斯山麓至波斯湾头的平原称胡齐斯坦。

The same Chinese text after the processing segmenting has become much more clear:

尔 兼德 高地 ， 北部 有 厄尔布兹 山脉 ， 德马万德峰 海拔 5670 米 ， 为 伊朗 最高 峰 。 西部 和 西南部 是 宽阔 的 扎 格罗斯 山山 系 ， 约占 国土 面积 一半 。 中部 为 干燥 的 盆地 ， 形成 许多 沙漠 ， 有 卡维尔 荒漠 与 卢特 荒漠 ， 平均 海拔 1，000余 米 。 仅 西南部 波斯湾 沿岸 与 北部 里海 沿岸 有 小 面积 的 冲击 平原 。 西南部 扎 格罗斯 山麓 至 波斯湾 头 的 平原 称 胡齐斯坦 。

As we can see, the boundaries of Chinese words, consisting of several characters, are clearly marked.

At the second stage, using the method of regular expressions, we pulled the group of the received characters in a vertical chain, and then translated it with an automatic translator. In the third stage, we chose a combination, satisfying the requirements of the terms. Particular attention was paid to extract terms from titles chapters and subchapters section “Afghanistan” representing the ontology information. As a result the terms were broken up into meaningful groups to compile a thesaurus of the Chinese Afghan Studies.

The section “Afghanistan” of the Chinese online encyclopedia Baidu were chosen by us as the object of investigation. Baidu is online encyclopedia in Chinese, which develops and supports the Chinese search engine Baidu. As well as Baidu itself, the encyclopedia is censored in accordance with Chinese government regulations. On June 2013 Baidu encyclopedia contained more than 6.2 million articles (more than English and German Wikipedia together) and had more than 3.2 million of participants.

Our work was divided into several stages:

1. selection of raw texts about Afghanistan in Chinese,
2. using the word processing program for automatic annotation of the text and isolation of terminological phrases,
3. updating the terminology.

The ontology, the geographical names of Afghanistan in Chinese transcription, ethnonyms peoples of Afghanistan and Central Asia [6], the names of political figures of Afghanistan in the Chinese transcription, the terms of political geography, the names of international organizations [7], Islamic concepts in Chinese, Arabisms and Farsisms in Chinese transcription became the object of special interest for our research. All these demonstrates the need for the development and introduction of special courses on teaching students how to work with the tools of computer NLP instruments.

References

1. Масс-медиа КНР в условиях глобализации // СИСП. 2012. №9. С.79.
2. Международное сотрудничество в терминологических исследованиях: Сб. Статей / Под науч. ред. К.К. Васильевой, Чжен Шупу. -Чита: Поиск, 2010.
3. Мишанкина Н. А. Базы данных в лингвистических исследованиях // Вопросы лексикографии. 2013. №1. С.25-33.
4. Нагель О. В. Корпусная лингвистика и ее использование в компьютеризированном языковом обучении // Язык и культура. 2008. №4. С.53-59.
5. Очиров О.Р. Лингвистические проблемы экономической терминологии современного китайского языка//Ученые записки Забайкальского государственного гуманитарно-педагогического университета им. Н.Г. Чернышевского. -2009. -№ 3. -С. 138-142.
6. Очиров О.Р. Становление китайского терминоведения: традиции и современность // Вестник Российского университета дружбы народов. Серия: Лингвистика. 2013. № 4. С. 116-125.
7. Очиров О.Р. Терминология современного китайского языка // Ученые записки Забайкальского государственного гуманитарно-педагогического университета им. Н.Г. Чернышевского. -2009. -№ 3. -С. 236-238.
8. Худякова О. С. Уровни ориентирующего воздействия специфических языковых структур и единиц в китайскоязычной блогосфере // Научный диалог. 2012. №3. С.138-160.
9. Чешуин С. А. Совершенствование профессиональной подготовки специалистов по лингвистике и межкультурной коммуникации на основе применения локальных вычислительных сетей // Армия и общество. 2009. №2. С.82-88.
10. Яцко В.А. Алгоритмы и программы автоматической обработки текста // Вестник ИГЛУ. 2012. №17.
11. Roger Levy and Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank?. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics — Volume 1 (ACL '03), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 439-446.
12. Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning. "A Conditional Random Field Word Segmenter." In Fourth SIGHAN Workshop on Chinese Language Processing. 2005.
13. Pi-Chuan Chang, Michel Gally and Christopher Manning. "Optimizing Chinese Word Segmentation for Machine Translation Performance" In ACL 2008 Third Workshop on Statistical Machine Translation.

Извлечение терминологии из энциклопедии Baidu

Булат Фаткулин[1,2]

[1] Южно-Уральский Государственный Университет, Челябинск, Россия
[2] Челябинский Государственный Университет, Челябинск, Россия
bfatkulin@gmail.com

Аннотация В статье рассматривается применение лингвистических технологий для преподавания ориенталистики в системе высшего образования Российской Федерации. Исследование посвящено методам извлечения терминологических единиц с использованием текстов об Афганистане на китайском языке. Приводятся решения для интенсивного автореферирования китайских текстов и предложены методы обучения студентов работе со вспомогательным программным обеспечением. Для сегментирования и извлечения именованных сущностей из текста на китайском языке использован пакет Stanford NLP.

Ключевые слова: извлечение терминологии, ориенталистика, обработка естественного языка.