**Выделение двуязычных терминов и терминологических словосочетаний на основе параллельного англо-русского корпуса по языкам программирования**

***Кощеева С.С.***

Санкт-Петербургский государственный университет (Санкт-Петербург), Россия

swetik-1995@mail.ru

**Аннотация**

В настоящее время параллельные корпусы текстов широко используются для выделения терминов и терминологических словосочетаний. В корпусной лингвистике были выработаны различные методы для автоматического извлечения однословных и многословных терминов. Данное исследование направлено на изучение эффективности статистического и лингвистического методов выделения терминов на материале параллельного англо-русского корпуса текстов по языкам программирования.

**Ключевые слова:** параллельные корпусы, термины, терминологические словосочетания, ключевые слова

**Abstract**

In recent years parallel corpora have been widely used for term extraction. Various methods of extracting single terms as well as multiword terms have been developed in corpus linguistics. This research aims at studying and comparing statistical and linguistic methods of term extraction, based on the parallel English-Russian corpus of programming languages.

**Keywords:** parallel corpora, terms, multiword terms, keywords

**1. Введение**

Параллельные корпусы текстов широко используются в корпусной лингвистике для выделения терминов и терминологических словосочетаний, которые затем можно применять в качестве основы для машинного перевода, создания терминологических словарей и тезаурусов предметной области, а также для верификации уже зафиксированных значений терминов и оценки адекватности их перевода.

*Цель* данного исследования заключается в том, чтобы на основе параллельного корпуса по языкам программирования изучить и сравнить различные методы выделения терминов и терминологических сочетаний.

**2. Извлечение терминов и терминологических словосочетаний из параллельного корпуса**

В соответствии с целью данного исследования был создан параллельный корпус, в который вошли тексты по программированию на русском и английском языках (источник: https://msdn.microsoft.com/library - официальный сайт документации Miсrosoft). Для выравнивания текстов использовалась простая утилита TextAligner. Затем с помощью корпусного менеджера SketchEngine был создан параллельный корпус общим объёмом около 260 000 токенов. В результате выравнивания текстов на уровне слов с помощью модуля Anymalign, написанного на языке программирования Python (<https://anymalign.limsi.fr/>), был получен двуязычный англо-русский словарь объёмом 11175 слов.

Выделение однословных терминов осуществлялось на основе метода ключевых слов. Для решения этой задачи использовался модуль Keywords корпусного менеджера SketchEngine, который определяет, является ли слово ключевым на основе его частоты в специальном корпусе по отношению к частоте данного слова в общеязыковом корпусе. Извлекались ключевые слова в форме лемм. С помощью созданного нами двуязычного англо-русского словаря ключевые слова, выделенные в английском языке, были переведены на русский язык. Если эти переводные эквиваленты обнаруживались среди ключевых слов, выделенных в русском языке, данные слова принимались за двуязычные термины. Такой алгоритм позволил выделить 481 релевантных переводных эквивалентов (11% от общего числа лемм), например, *type – тип, class – класс, method – метод, code – код, property – свойство*.

На этапе выделения терминологических словосочетаний мы ограничились биграммами и исследовали статистический и лингвистический методы их выделения.

Выделение двуязычных биграмм статистическим методом осуществлялось с помощью модуля Terms в системе SketchEngine, основанного на формуле меры относительной частоты. Для установления соответствия между выделенными биграммами на языке программирования C# был написан программный код, с помощью которого выделенные английские конструкции преобразуются в форму лемм, леммы с помощью созданного двуязычного словаря переводятся на русский язык и получившиеся формы сравниваются с выделенными и переведёнными конструкциями в русском языке.

Данный алгоритм позволил установить соответствие между 124 биграммами русского и английского языка. На основе словарей ABBYY Lingvo (Англо-русский словарь по вычислительной технике и программированию и Англо-русский научно-технический словарь) была проведена оценка точности статистического метода, в результате которой лишь 30 биграмм оказались релевантными, что составляет около 24% точности. Релевантными двуязычными биграммами, например, являются: *generic class – универсальный класс , base class – базовый класс, class library – библиотека классов, console application – консольное приложение*.

В рамках исследования лингвистического метода были выработаны морфосинтаксические шаблоны терминологических биграмм для английского и русского языков и установлены соответствия между ними (см. Таблица 1).

Таблица 1

**Типы соответствий между морфосинтаксическими шаблонами**

|  |  |  |
| --- | --- | --- |
| **Шаблоны английского языка** | **Шаблоны русского языка** | **Примеры** |
| Adj + Noun | Adj + Noun | generic type – универсальный тип |
| Noun1 + Noun2 | Adj + Noun2 | base type – базовый тип |
| Noun1 + Noun2 | Noun1 + Noun2(Gen) | implementing interface – реализация интерфейса |
| Noun1 + Noun2 | Noun1 + Noun2(Instr) | managing assemblies – управление сборками |
| Noun1 + Noun2 | Noun2 + Prep + Noun1 | method access –доступ к методам |
| Noun1 + of + Noun2 | Noun1 + Noun2(Gen) | type of delegate – тип делегата |
| Noun1 + Noun2 | Noun2 + Noun1(Gen) | code editor – редактор кода |
| Noun1 + Prep + Noun2 | Noun1 + Prep + Noun2 | access to store – доступ к хранилищу |

В таблице 2 отображены результаты установления соответствий между выделенными словосочетаниями в русском и английском языке.

Таблица 2

**Оценка релевантности словосочетаний, выделенных с помощью лингвистического метода**

|  |  |  |
| --- | --- | --- |
| **Тип соответствия (англ. – рус.)** | **Количество биграмм** | **Точность** |
| **Общее число** | **Релевантные** |
| Adj + Noun – Adj + Noun  | 404 | 73 | 18% |
| Noun1 + Noun2 – Noun1 + Noun2(Gen)  | 178 | 0 | 0% |
| Noun1 + Noun2 – Noun2 + Noun1Gen | 577 | 154 | 27% |
| Noun1 + of + Noun2 – Noun1 + Noun2(Gen)  | 131 | 14 | 11% |
| Noun1 + Noun2 – Noun1 + Noun2(Instr)  | 12 | 4 | 33% |
| Noun1 + Prep + Noun2 – Noun1 + Prep + Noun2  | 107 | 0 | 0% |
| Noun1 + Noun2 – Noun2 + Prep + Noun1  | 193 | 11 | 6% |
| Noun1 + Noun2 – Adj + Noun2  | 170 | 27 | 16% |

Примеры релевантных переводных эквивалентов: *generic class* – универсальный класс; *stack overflow* – переполнение стека; *instance of class* – экземпляр класса; *access control* – управление доступом; *data storage* – хранилище для данных; *helper function* – вспомогательная функция.

Средняя точность лингвистического метода в рамках разработанного нами алгоритма составляет, таким образом, 16 %.

**3. Выводы**

Анализ статистического и лингвистического метода выделения двуязычных терминологических биграмм показал, что ни один из методов не отличаются высокими показателями точности (24% для статистического метода и 16% для лингвистического метода). Однако использование морфосинтаксических шаблонов позволяет выделять более разнообразные типы терминологических словосочетаний, в отличие от статистического метода, который выдаёт очень ограниченное количество результатов.

Различие статистического и лингвистического метода заключается и в объёме предварительной подготовки и обработки текстов. Статистический метод используется, в основном, на необработанном корпусе текстов, поэтому не требует особой предварительной работы над текстами корпуса. Выделение терминологических биграмм на основе морфосинтаксических шаблонов требует морфологической и лемматической разметки корпуса текстов, что повышает трудоёмкость данного метода.

Таким образом, статистический и лингвистический методы не отличаются высокой точностью, но тем не менее могут быть использованы для выделения терминологических словосочетаний. Предполагается, что совместное использование статистического и лингвистического методов (смешанный метод) может значительно повысить качество выделения двуязычных терминов и терминологических словосочетаний в параллельных корпусах текстов.

**Список литературы**

1. Браславский, П., Соколов, Е. Сравнение пяти методов извлечения терминов произвольной длины. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2008». 2008. С. 67–74.
2. Guinovart X., Simoes A. Parallel corpus-based bilingual terminology extraction // Proceedings of the 8th International Conference on Terminology and Artificial Intelligence. – Toulouse. 2009.
3. Kilgariff A. Finding Terms in Corpora for Many Languages with the Sketch Engine // Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. 2014. P. 53–56.