БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
КАФЕДРА ИНФОРМАЦИОННЫХ СИСТЕМ УПРАВЛЕНИЯ

Дипломная работа

Исаченко Дмитрия Александровича

CROSS-LANGUAGE ФУНКЦИОНАЛЬНОСТЬ АВТОМАТИЧЕСКОГО ПОИСКА В СЕТИ INTERNET РЕЛЕВАНТНЫХ ДОКУМЕНТОВ

Руководитель
Совпель Игорь Васильевич,
доктор технических наук, профессор
Рецензент
Липницкий Станислав Феликсович
доктор технических наук,
гл.н.с. ГНУ «ОИПИ НАН БЕЛАРУСИ»

Актуальность темы

- Информации в интернете много, но она представлена не на всех языках;
- Язык извлечённой поисковыми системами информации такой же, как и язык исходного поискового запроса.

Существующие решения

Google

- Возможности:
  1. Предоставляет большой список возможностей для поиска: поиск по карте, картинкам, видео…
  2. Поддерживает cross-language функциональности при поиске.

- Чего не хватает:
  1. Увеличение максимальной длины запроса(текущее ограничение - 2048 сивмолов);
  2. Поддержки в качестве входных для поиска данных веб-страницы, текстового документа(PDF, TXT).

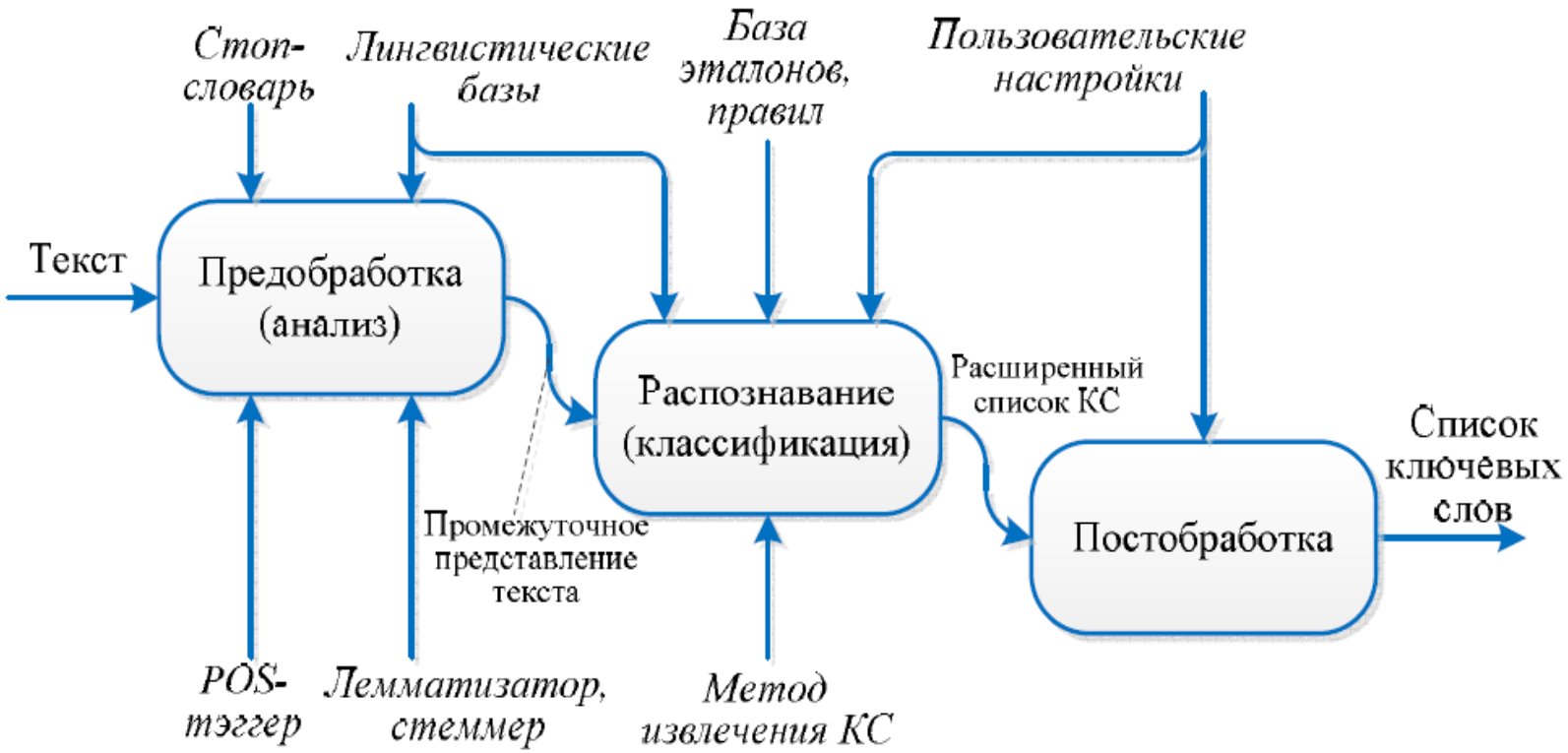
Цели и задачи работы

Целью данной дипломной работы является разработка мобильного приложение, обеспечивающего поиск в сети интернет по заданному текстовому документу/веб-странице релевантных документов, в том числе представленных на языке отличном от языка входных данных.

Для достижения поставленной цели необходимо решить следующие **задачи**:

1. Разработать алгоритм составления ПОД;
2. Разработать алгоритм машинного перевода;
3. Разработать структурно-функциональную схему системы поиска релевантных документов;
4. Реализовать мобильное приложение.

Составление поскового образа документа



Составление поскового образа документа

Сервисы/инструменты для извлечения ключевой информации из текста:

1. OpenCalais (от Thomson Reuters);
2. IBM's Watson Natural Language Understanding Service;
3. Yahoo Content Analysis;
4. Mining Cloud (ранее Text Analytics);
5. Stanford's Core NLP Suite ;
6. Natural Language Toolkit;
7. Apache OpenNLP.

IBM's Watson Natural Language Understanding Service

200 OK

Headers >

Response body


```
{
    "keywords": [{
        "text": "clinical trials",
        "relevance": 0.990117
    }, {
        "text": "treatment",
        "relevance": 0.842602
    }, {
        "text": "cancer",
        "relevance": 0.766056
    }, {
        "text": "research studies",
        "relevance": 0.744387
    }, {
        "text": "hormone therapy",
        "relevance": 0.736107
    }, {
        "text": "radiation therapy",
        "relevance": 0.733394
    }],
    "language": "en"
}
```


Список извлечённых ключевых слов

200 OK

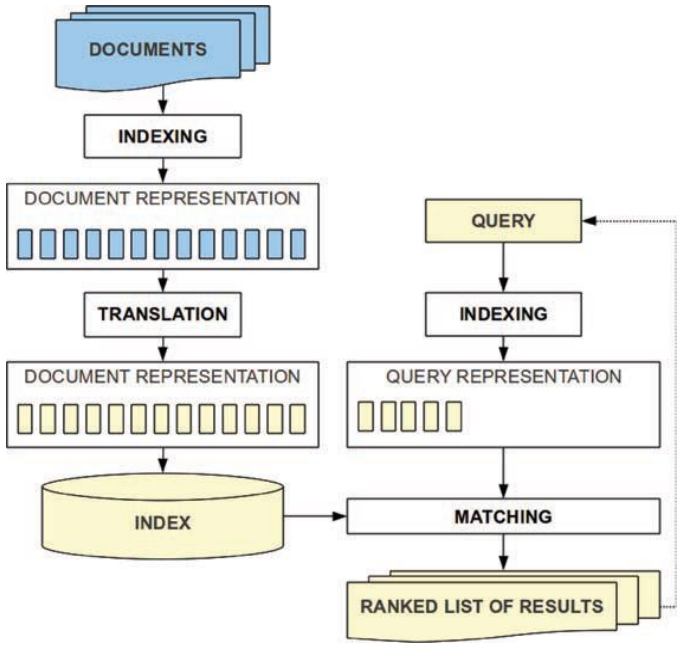
Headers >

Response body

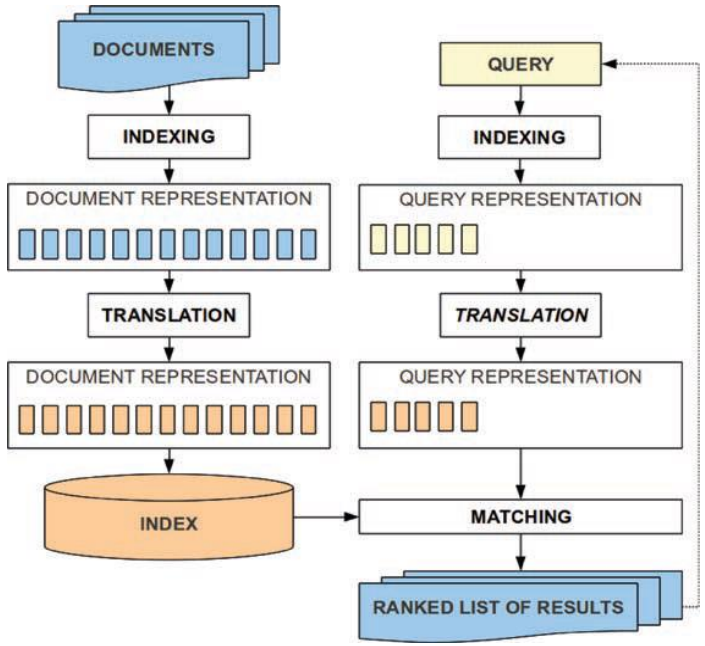

```
{
    "semantic_roles": [{
        "subject": {
            "text": "The types of treatment"
        },
        "sentence": "The types of treatment that you have will depend on the type of cancer you have and how advanced it is.",
        "object": {
            "text": "on the type of cancer you have"
        },
        "action": {
            "verb": {
                "text": "depend",
                "tense": "future"
            },
            "text": "will depend",
            "normalized": "will depend"
        }
    }, {
```


Выделение контекста для ключевого слова

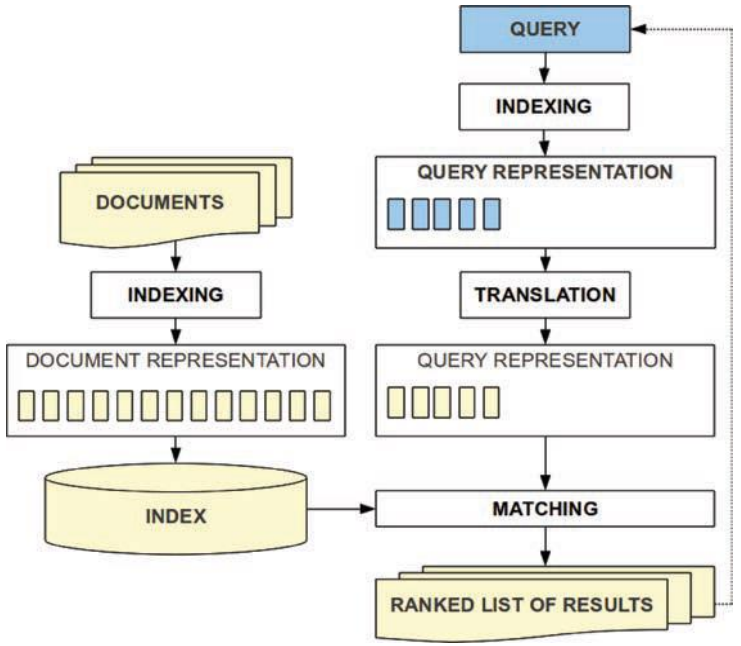
Концепции машинного перевода



Перевод документов



Перевод документов и ПОЗа



Перевод ПОЗа

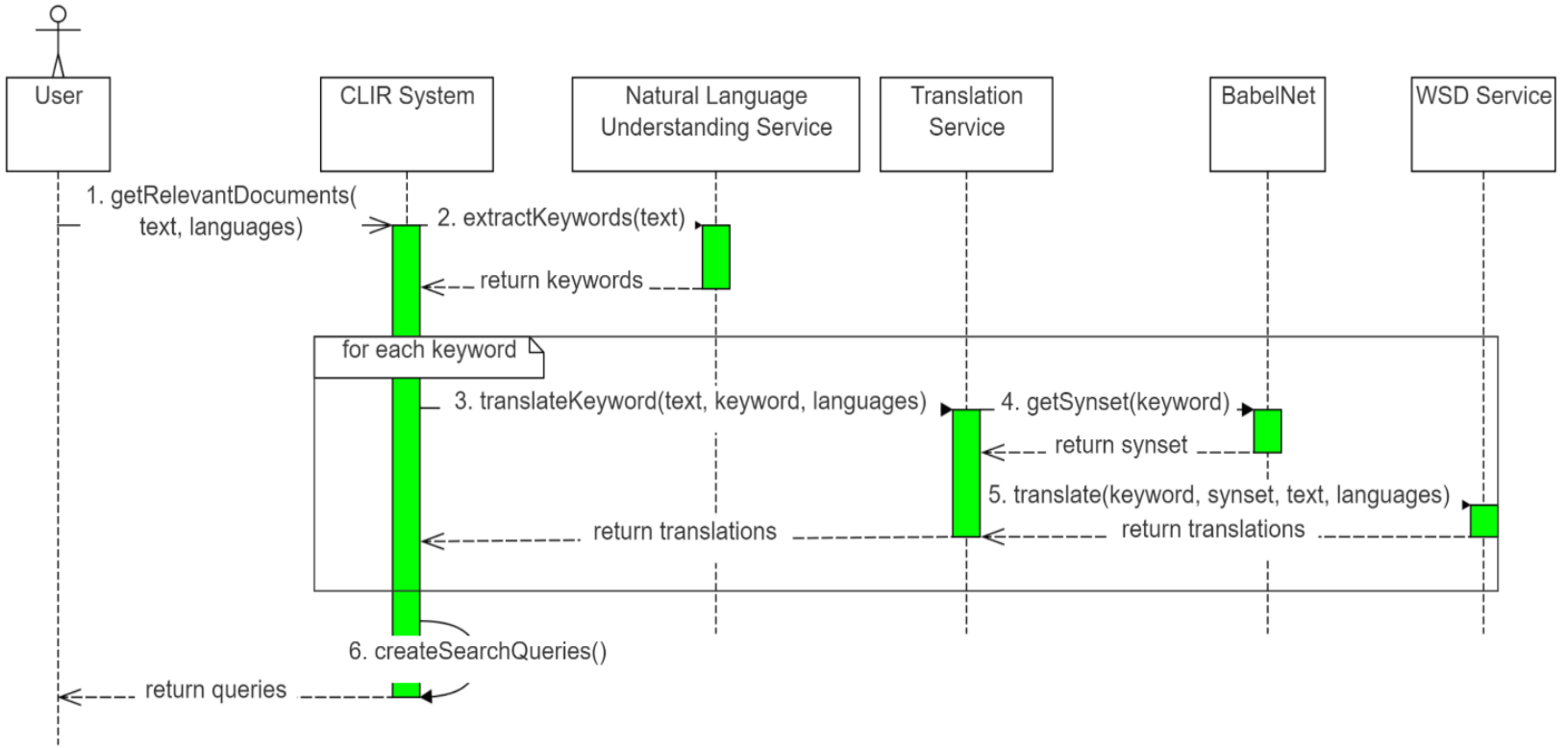
Разрешение лексической многозначности при переводе

Синсет и соответствующие ему значения	Русский перевод	Французский перевод
treatment, intervention **Provided to improve a situation (especially medical procedures or applications that are intended to relieve disease or injury).** **Medical care for an disease or injury.** **A treatment or cure is applied after a medical problem has already started.**	лечение	traitement
treatment, handling **The management of someone or something.**	обращение	traitement
A manner of dealing with something artistically	трактовка	traitement
discourse, treatment, discussion, speech **An extended communication dealing with some particular topic.**	Дискурс, доклад, лекция	discours, discours politique, discours public

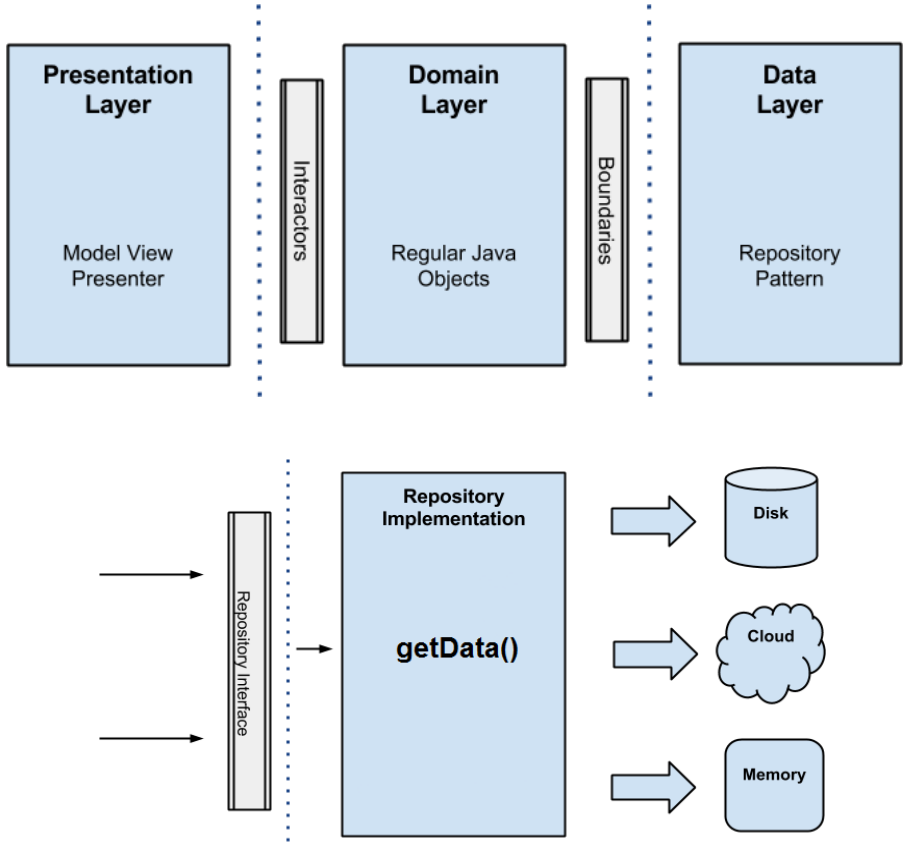
Набор синсетов для “treatment”

1. Разрешение лексической многозначности происходит по алгоритму Леска
2. В качестве внешнего источника знаний используется многоязычный электронный тезаурус BabelNet.

Диаграмма последовательности поиска релевантных документов



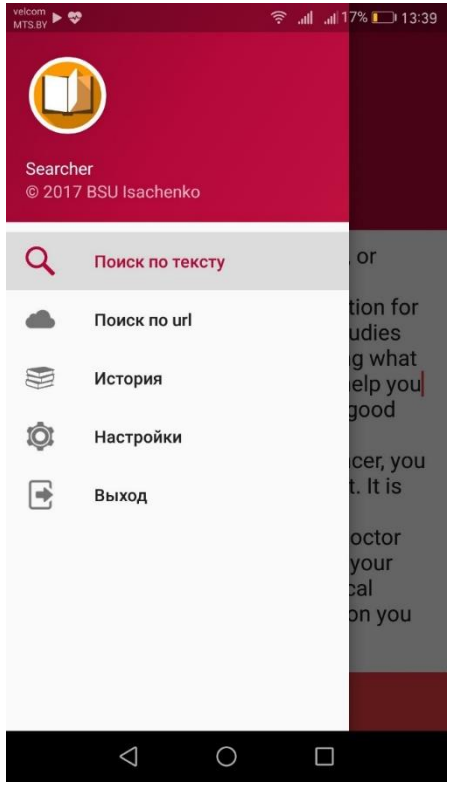
Архитектура приложения

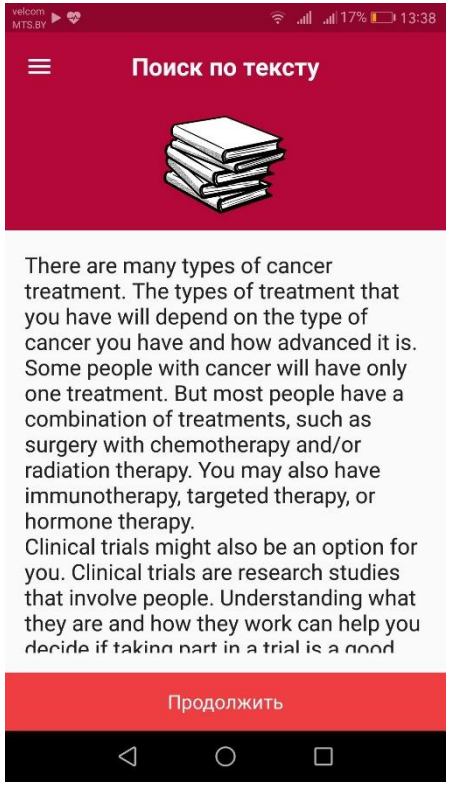


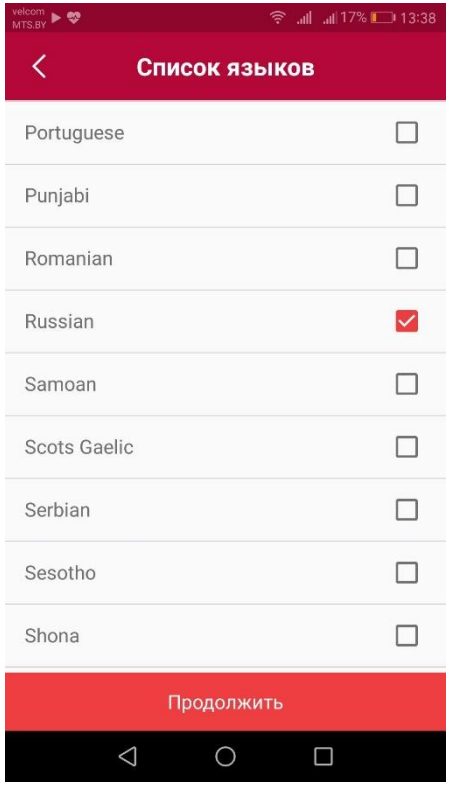
Преимущества архитектуры:

- Независимость от внешних сервисов, с которыми взаимодействует приложение;
- Независимость от фреймворков;
- Независимость от Баз Данных;
- Независимость от UI;
- Простота написания тестов.

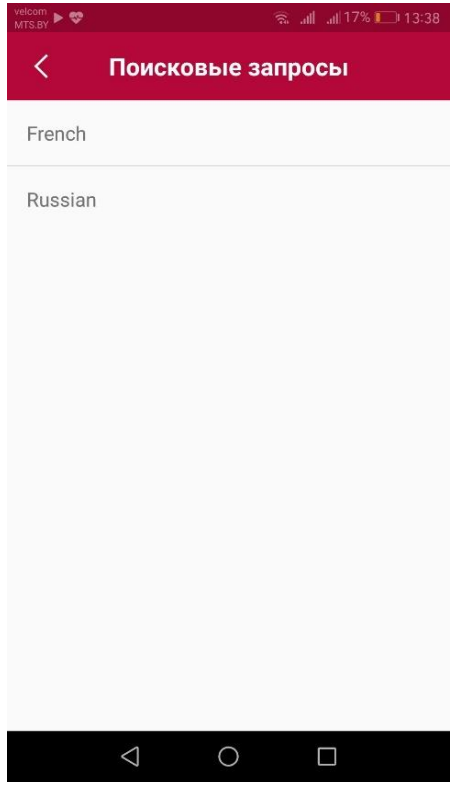
Демонстрация работы приложения

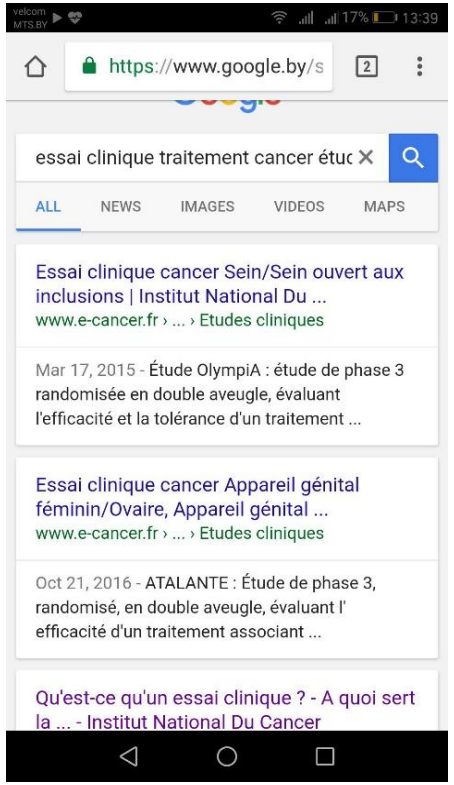


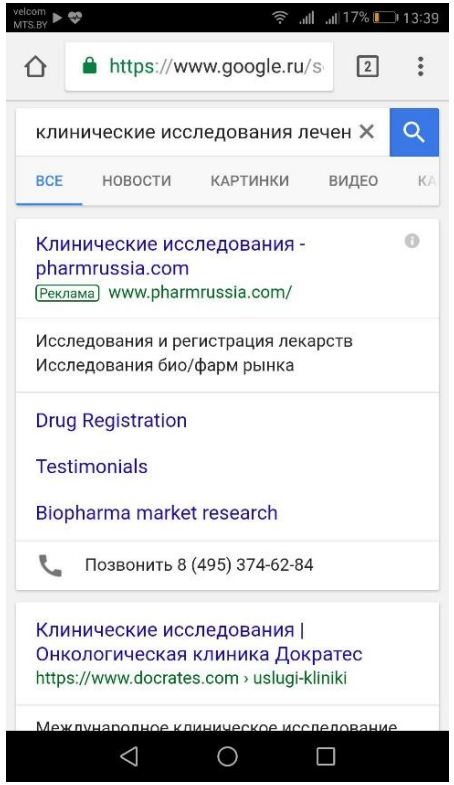




Демонстрация работы приложения







Результаты

- Выполнен анализ сервисов, предоставляющих возможность извлечения ключевой информации из текста;
- Исследованы концепции машинного перевода, а так же алгоритмы, применяющиеся для разрешения лексической многозначности при переводе слов;
- Построена структурно-функциональная схема системы поиска релевантных документов;
- Разработано мобильное приложение под ОС Android, обеспечивающего поиск в сети интернет по заданному текстовому документу/веб-странице релевантных документов, в том числе представленных на языке отличном от языка входных данных.

Спасибо за внимание!