Открытая программная платформа UIMA-Ext для разработки приложений анализа русскоязычных текстов

Гареев Ринат, Иванов Владимир,
Высшая школа ИТИС
Казанского федерального университета

План доклада

- Мотивация
- “UIMA-Ext”
- UIMA и UIMA Async Scaleout
- Модули UIMA-Ext
- Textocat

Проблема

Существует множество программных инструментов для обработки естественного языка (NLP).

http://nlpub.ru/Обработка_текста

Сложности разработчиков конечных приложений:

- интеграция инструментов между собой,
- доступность инструмента (free vs $), открытость исходного кода, актуальность,
- расширяемость, в т.ч., адаптируемость для русскоязычного входа,
- масштабируемость.

Примеры

AOT.ru

ЭТАП-3

Apache OpenNLP, LingPipe

NLTK, GATE, UIMA

Mystem

Abbyy Compreno, RCO

Наш проект - “UIMA-Ext”

Наработки научной группы (рук. - проф. Соловьев В.Д., Казанский университет):

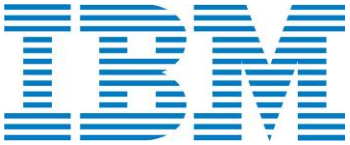
- базовые инструменты NLP для русского языка,
- в ходе выполнения ряда проектов (HP Labs, “Династия”, РФФИ),
- на единой платформе Apache UIMA.

Src: https://github.com/CLLKazan/UIMA-Ext

Docs (in progress): http://cllkazan.github.io/UIMA-Ext/

Apache Software License v.2

Apache UIMA



“*Unstructured Information Management Architecture*”

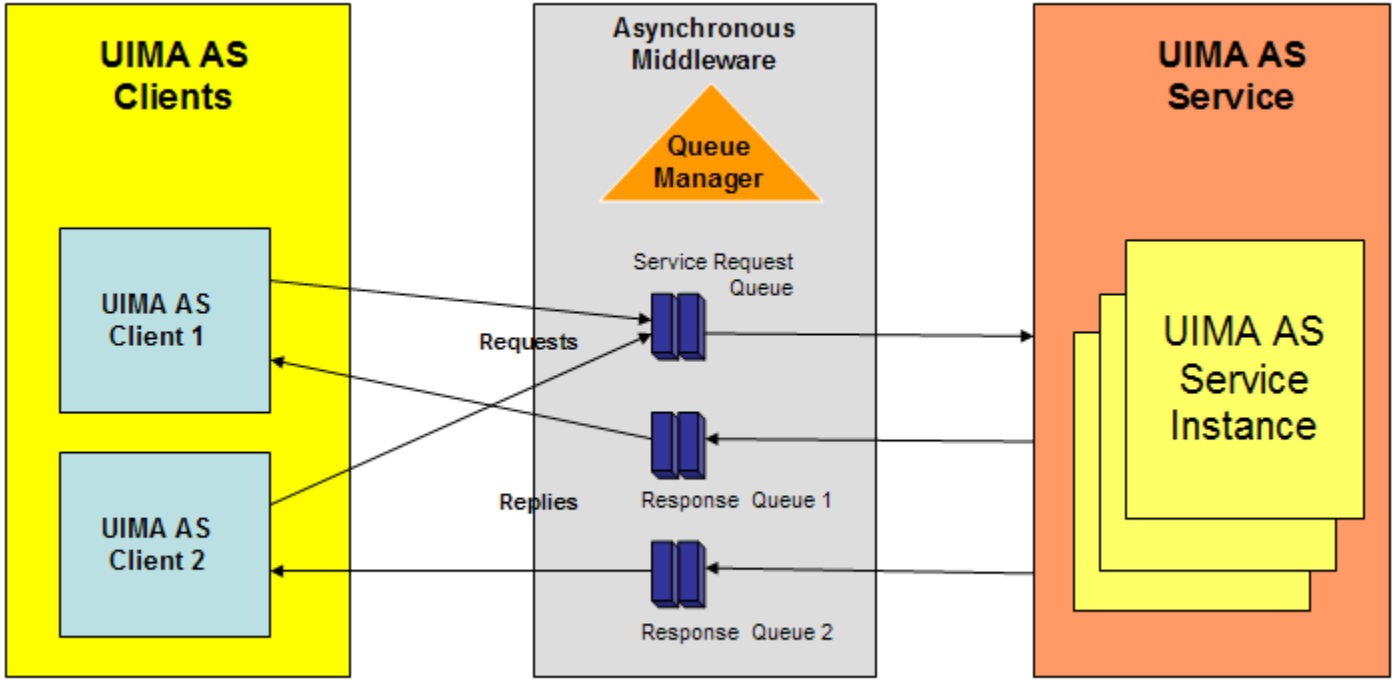
Архитектура, которая специфицирует интерфейсы:

- Common Analysis System (CAS) - представление текста и графа аннотаций над этим текстом
- обработчики текста (“annotator”),
- разделяемые ресурсы (“shared resource”), н-р, для словарей, “обученных” моделей классификаторов и т.п.

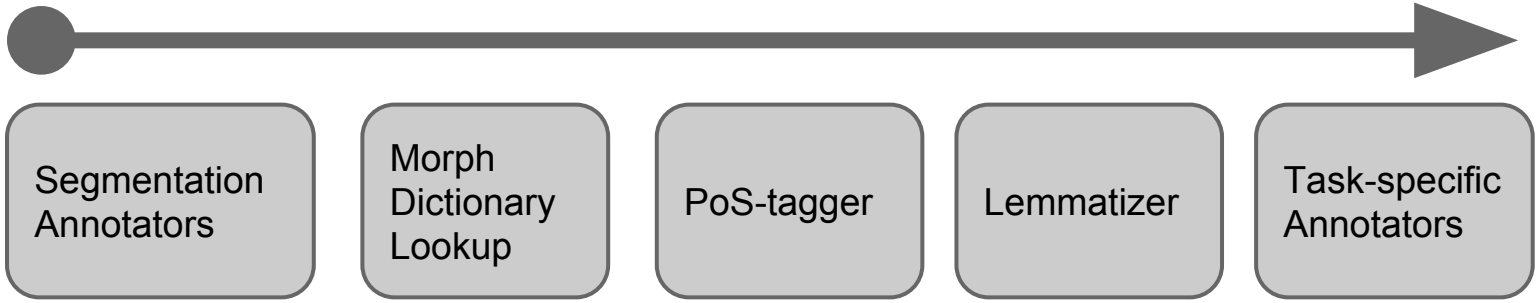
Преимущества UIMA

+ разделение алгоритмов, структур данных и инфраструктуры “развертывания” (deployment), спецификация OASIS UIMA
+ гибкая смена компонент в “конвейере” (pipeline)
+ удобное тестирование / экспериментирование / развертывание

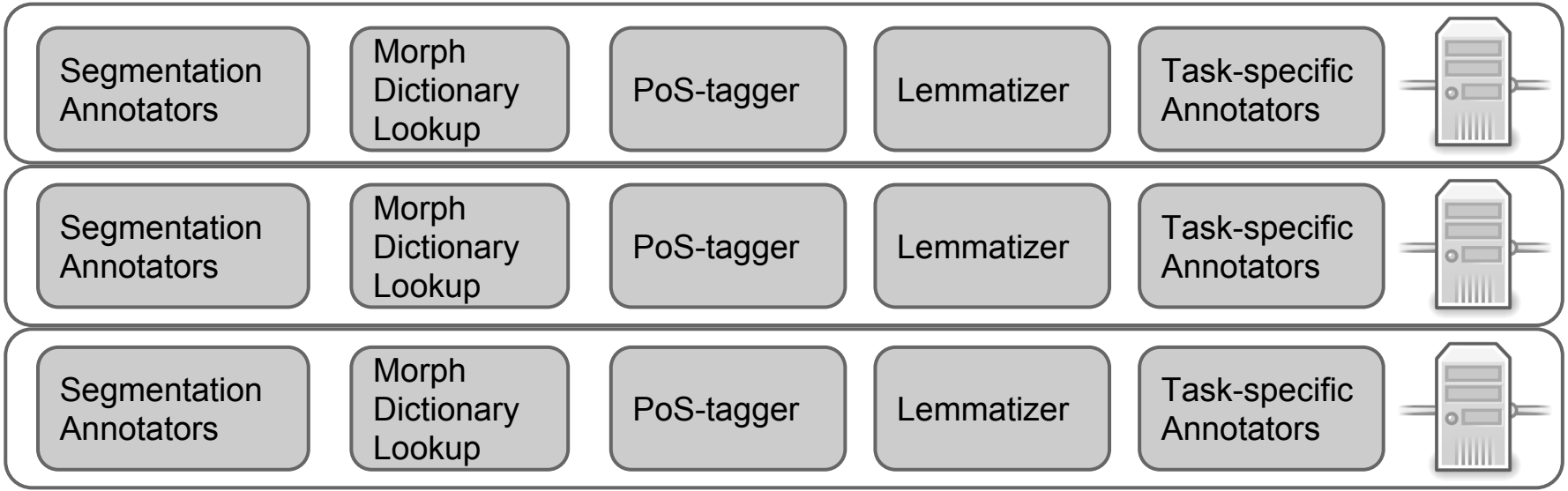
UIMA Async Scaleout



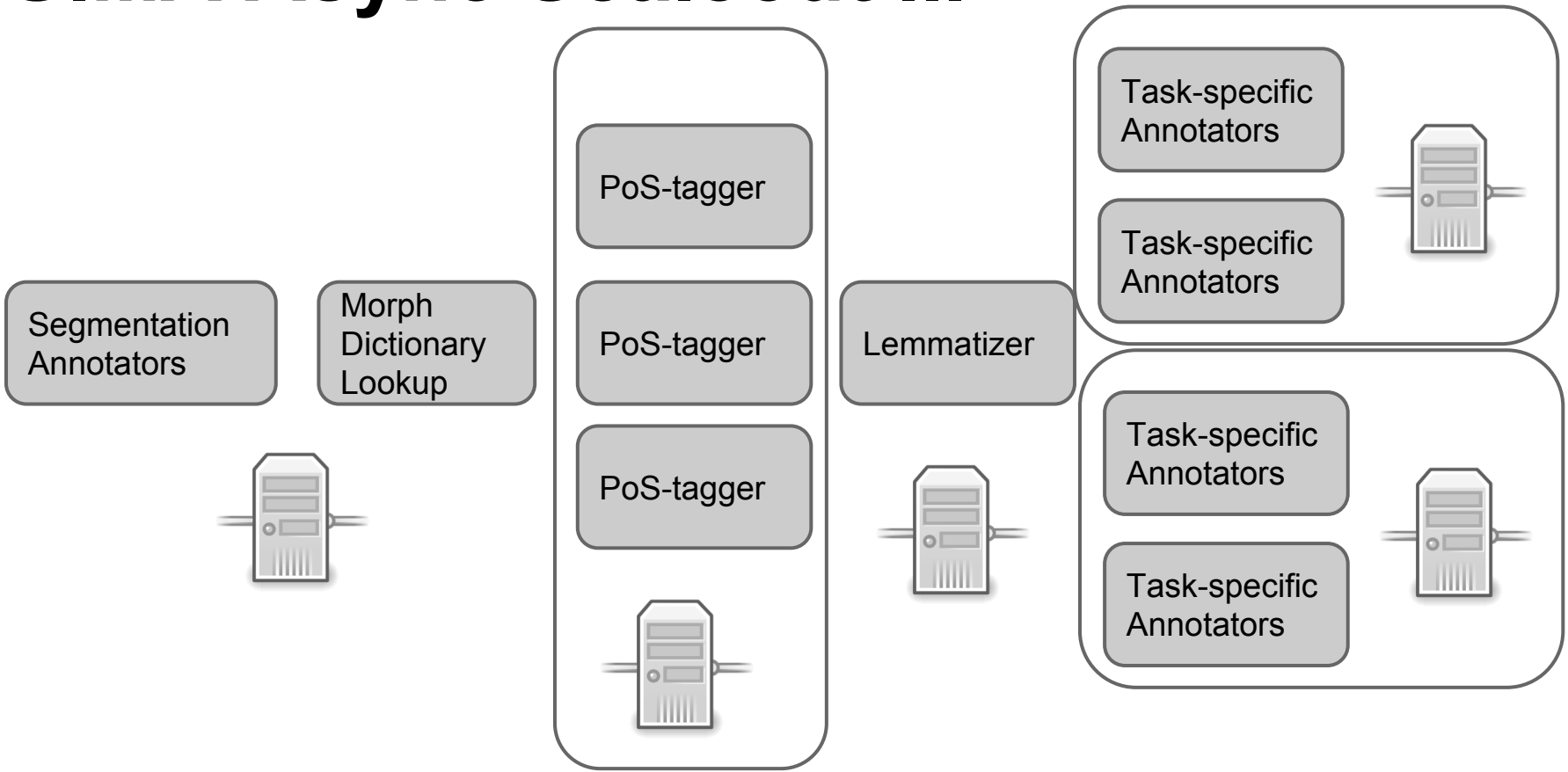
UIMA Async Scaleout - пример конвейера



UIMA Async Scaleout II



UIMA Async Scaleout III



IBM Watson

“...We used UIMA-AS to scale Watson out over 2500 compute cores...”,

Ferrucci, David, et al. "Building Watson: An overview of the DeepQA project." AI magazine 31.3 (2010): 59-79.

Модули UIMA-Ext

- Сегментация текста
- Словарная морфология
- PoS-tagging
- Лемматизация
- Чанкинг (NP recognition)
- Синтаксический анализатор
- Универсальный модуль оценивания качества
- Инструменты для создания корпусов

Морфологический анализ

- интерфейс для работы со словарями UIMA.Ext.Dictionary.API
- готовая реализация для словаря

- предсказание характеристик для несловарных форм

Снятие омонимии (PoS-tagging)

Национальный корпус русского языка (1M)

~500 “тэгов”

Реализация	**Accuracy**
Только словарь	80.5%
HunPos (+словарь)	90.6%
Stanford PoS-tagger (без словаря)	90.7%
OpenNLP MaxEnt (+словарь)	92.1%
Tiered Conditional Random Fields (+словарь)	93.3%

Лемматизация

Приведение к словарной форме

“его” => “его”, “его” => “он”

“для” => “длить”, “для” => “для”

UIMA.Ext.Lemmatizer.API

UIMA.Ext.Lemmatizer.OpenCorpora

Выделение базовых именных групп

Пример:

[[Динамика курса рубля]], наблюдаемая [[в настоящее время]], создает [[предпосылки]] [[к возникновению рисков]] [[для финансовой стабильности]] и [[формированию устойчивых девальвационных и инфляционных ожиданий]], заявил [[Банк России]] [[в четверг]]

Реализация:

контекстно-свободная грамматика,

Scala Parser Combinators

Синтаксический анализ

На основе MSTParser

Обучен на подкорпусе СинТагРус

84% UAS (Unlabelled Attachment Score)

Инструменты для разработки корпусов

- интеграция с Brat Rapid Annotation Tool (http://brat.nlplab.org)
- контроль качества: подсчет коэффициентов согласованности между разметчиками
- (в процессе разработки) Rapid Manual Annotation (RMA) Tool:
  - один back-end (REST API) для разных интерфейсов (web, mobile, AMTurk,...)
  - для декомпозиции сложных задач лингвистической разметки на несколько элементарных

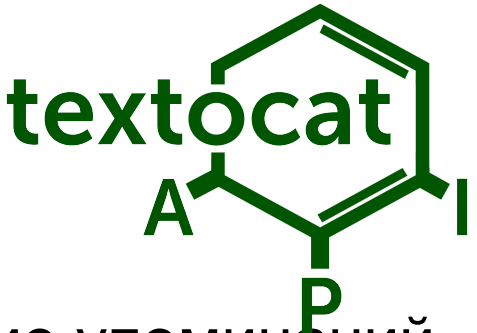
Публикации

1. Gareev R., Tkachenko M., Solovyev V., Simanovsky A., Ivanov V. Introducing baselines for Russian named entity recognition // CICLING 2013. – Springer, 2013. – С. 329-342.
2. Gareev R., Ivanov V. A comparative evaluation of statistical Part-of-Speech taggers for Russian // RUSSIR 2014 YSC. — Communications in Computer and Information Science,— Springer, 2014. *В печати*...
3. Ivanov V. Extracting Frame-Like Structures from Google Books NGram Dataset // Human-Inspired Computing and Its Applications. – Springer, 2014. – С. 18-27.
4. Solovyev V., Ivanov V. Dictionary-Based Problem Phrase Extraction from User Reviews // TSD 2014. – Springer, 2014. – С. 225-232.
5. Ivanov V., Tutubalina E. Clause-based approach to extracting problem phrases from user reviews of products // AIST 2014. – Springer, 2014. – С. 229-236.

Применения

Textocat - облачный сервис для обработки русскоязычных текстов

http://textocat.com/demo



Текущий функционал: распознавание упоминаний персон, организаций, населенных пунктов, денежных и временных выражений

Спасибо за внимание!