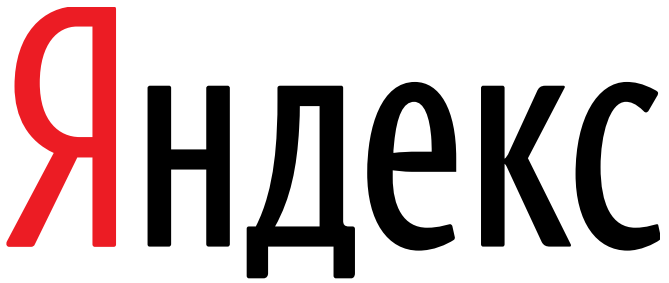
Яндекс

Илья Сегалович и лингвистика в Яндексе

Михаил Маслов, Алексей Зобнин

Лингвистические проекты Ильи

- Участие Яндекса в проекте НКРЯ
- Морфологический анализ и синтез для неизвестных слов
- Снятие морфологической омонимии
- Практическая транскрипция имен собственных
- Расстановка ударений и определение размера стиха

Морфологический анализ и синтез для неизвестных слов

Морфологический анализ и синтез

- 1995-96: поисковая система и конкорданс для ЭНИ «Грибоедов», «Информ-Норматив»
- Обработка несловарных слов
  - Поиск в словаре слова с максимальной длиной совпавшего «хвоста»
  - Генерация гипотез моделей словоизменения
  - Попытка выбрать лучшую модель по статистике корпуса ЭНИ

(И. Сегалович, М. Маслов. Диалог-1998)

Морфологический анализ и синтез

- 1995-96: поисковая система и конкорданс для ЭНИ «Грибоедов», «Информ-Норматив»
- Обработка несловарных слов
  - Поиск в словаре слова с максимальной длиной совпавшего «хвоста»
  - Генерация гипотез моделей словоизменения
  - Попытка выбрать лучшую модель по статистике корпуса ЭНИ

(И. Сегалович, М. Маслов. Диалог-1998)

Морфологический анализ и синтез

глокая	{глокать? глокий?}
куздра	{куздра?}
штеко	{штекий? штеко?}
будланула	{будланул? будланула? будлануть?}
бокра	{бокр? бокра? бокрый?}
и	{и}
кудрячит	{кудрячит? кудрячита? кудрячитый? кудрячить?}
бокренка	{бокренк? бокренка? бокренок?}

Национальный корпус русского языка

Создание НКРЯ

1999 **Яндексу нужен размеченный корпус**

2001 **Яндекс + Ruscorpora**

2004 **Запуск Ruscorpora**

Размер НКРЯ

- 2004 год:
  - корпус со снятой омонимией: ≈700 тыс. словоуп.
  - основной корпус: ≈11 млн словоуп.
- Сейчас:
  - корпус со снятой омонимией : ≈6 млн. словоуп.
  - основной корпус: ≈250 млн. словоуп.
  - всего ≈500 млн. словоуп.

Снятие морфологической омонимии

Снятие омонимии: постановка задачи

- Цели
  - поисковый конкорданс (для поиска по Вебу не нужен)
  - увеличение точности поиска
  - уменьшение объема индекса

- Задача
  - построить алгоритм выбора правильного текста леммы с высокой точностью и высокой производительностью (10^3 – 10^4 слов в секунду)

Снятие омонимии: подходы

- Детерминированные методы (ЭТАП, Диалинг)
  - Недостаточная точность
  - Небольшая производительность
- «Вероятностные» методы (HMM и т.п.)
  - Нужен размеченный корпус

Идеи метода

На входе: результат морфоразбора (mystem)

```
начала  {начинать=V,пе=прош,ед,изъяв,жен,сов}
        {начало=S,сред,неод,(им,мн|род,ед|вин,мн)}
```

Базовая единица словаря контекстов:

```
<омоним, элемент_контекста, лемма>
```

Вариант «очевидного» представления контекста

```
V,пе,…|S,сред,неод,…    <.>+2    начало    p1
V,пе,…|S,сред,неод,…    <.>+2    начинать  p2
```

(Ю. Зеленков, И. Сегалович, В. Титов, Диалог-2005)

Ключевая идея метода

Нормализующие подстановки!

```
начала =>  ала(1о|3инать)
```

Не только `начала`, но и

```
        гнала
        знала
      ужинала
   вспоминала
      окунала
            …
```

(Ю. Зеленков, И. Сегалович, В. Титов, Диалог-2005)

Пример из словаря контекстов

```
ала(1о|Зинать)    <.>+2    1о       0.67
ала(1о|Зинать)    <.>+2    Зинать   0.33
```

(Ю. Зеленков, И. Сегалович, В. Титов, Диалог-2005)

Используемые корпуса

Корпус со снятой омонимией НКРЯ-2004	700 тыс. словоуп.
Веб-корпус «трудных омонимов»	2 млн. словоуп.

(Ю. Зеленков, И. Сегалович, В. Титов, Диалог-2005)

Сравнение методов снятия омонимии*

	Точность	Улучшение baseline
Наиболее частотная лемма (baseline)	89.5%	
Граммемы	90.5%	15%
Нормализующие подстановки на НКРЯ—2004	93.5%	35%
Нормализующие подстановки на веб-корпусе	96.5%	65%
Нормализующие подстановки на НКРЯ—2014	97%	70%

* Зеленков Ю.Г.

Практическая транскрипция собственной лексики

Постановка задачи

- Личные имена, географические названия, названия групп, брендов и т. д.
- Применение в 2009
  - запросы к «большому» поиску
  - запросы Я.Музыке и т.п.
  - поиск по адресной книге книге в Я.Почте

Постановка задачи

- Задача
  - Построить универсальный вероятностный метод автоматической транскрипции
- Данные
  - 100 тыс. билингв (из Википедии и т.п.) для 17 языков

Методы практической транскрипции (Ю. Зеленков)

- Триграммы
- Марковские модели
- Байесовский подход
- Метод сегментов
  - » Сегмент – группа рядом стоящих гласных или рядом стоящих согласных букв
- И т.п. (всего методов – штук 8)

Методы практической транскрипции (выбор Ильи)

- Триграммы
- Марковские модели
- Байесовский подход
- Метод сегментов!
  - » Сегмент – группа рядом стоящих гласных или рядом стоящих согласных букв
- И т.п. (всего методов – штук 8)

Практическая транскрипция: метод сегментов

- **Обучение**
  - Выравниваем билингвы по сегментам
  - Строим список пар-переводов сегментов:
    1. учитываем левый и правый контекст
    2. Вычисляем вероятности

- **Применение**

Из всех вариантов переводов выбирается вариант с наибольшей суммой вероятностей

Выравнивание сегментов

Выравниванием побуквенно:

s	t	a	-	t	e	m	e	n	t
с	т	е	й	т	-	м	е	н	т

Выравнивание сегментов

Собираем сегменты, убираем пропуски в оригинале:

st	a	t	e	m	e	nt
ст	ей	т	-	м	е	нт

Выравнивание сегментов

Убираем пропуски в «переводе»:

st	**a**	**tem**	**e**	**nt**
ст	**ей**	**тм**	**е**	**нт**

Метод сегментов: пример

[r e n au lt] => [р е н о *] (язык fr)

```
   r          => р(0.95) *(0.03) рь(0.01) ж(0.01)
(^)r          => р(0.99) ж(0.01)
   r(e)       => р(0.98) ж(0.02)
   r(e n)     => р(0.95) ж(0.05)
```

Метод сегментов: пример

[r e n au lt] => [р е н о *] (язык fr)

```
     e          => e(0.49)*0.40 э(0.07) а(0.02) и(0.02)
  (r)e          => e(0.69)*0.25 э(0.03) а(0.02) и(0.01)
     e(n)       => e(0.90) э(0.06) а(0.02)*0.02
(^ r)e          => e(0.93) э(0.04) ё(0.02) ей(0.01)
     e(n au)    => e(1.00)
```

Метод сегментов: пример

[r e n au lt] => [p e н o *] (язык fr)

```
     n          => н(0.95) нь(0.04)*0.01
   (e)n         => н(0.95) нь(0.03)*0.02
     n(au)      => н(1.00)
(r e)n          => н(0.92) нь(0.07)*0.01
     n(au lt)   - отсутствует в выборке
```

Метод сегментов: пример

[r e n au lt] => [р е н о *] (язык fr)

```
     au        => o(0.54) ay(0.40) а(0.03) яу(0.02) оу(0.01)
  (n)au        => ay(0.54) o(0.38) ай(0.08)
     au(lt)    => o(0.96) ай(0.04)
(e n)au        => ay(0.50) o(0.50)
     au(lt $)  => o(0.96) ай(0.04)
```

Метод сегментов: пример

[r e n au lt] => [р е н о *] (язык fr)

lt => **льт(0.43)** лт(0.39)*0.16 лть(0.01) ль(0.01)

(au)lt => **λ(0.85)** ть(0.06) л(0.03) лт(0.03) льт(0.03)

lt($) => **λ(0.48) льт(0.33) лт(0.16)** л(0.02) ль(0.01)

(n au)lt => **λ(0.80) л(0.20)**

Обратная транскрипция на русский язык

- Проблема
  - Метод с вероятностями переводов работает плохо, точность ≈0.5
- Решение: яндексология
  - Яндексология всех вариатов работает тоже плохо
  - Выбираем 2-3 самых вероятных варианта
  - Применяем к ним яндексологию
  - Точность ≈0.9-0.95

Обратная транскрипция: примеры

```
милен демонжо     => Mylene Demongeot

марион котияр     => Marion Cotillard
марион котийяр    => Marion Cotillard
марион котийар    => Marion Cotillard
марион котилар    => Marion Cotilard

Эйяфьятлайокудль =>  Eyjafjallajokull
Эйяфьядлайёкюдль => Eyjafjyadlajokull
Эйяфьядлайёкудль => Eyjafjyadlajokull
```

Новое применение 2012-13 – карты всего мира по-русски (В. Титов)

Язык	Оригинал	Перевод
TUR	Ragıp Üner Caddesi Çocuk Parkı İç Yolları Ekmekçi Camii Çıkmazı	Рагып Унер Джаддеси Чоджук Паркы Ич Йоллары Экмекчи Джамии Чикмазы
FRE	Sente de l'Abreuvoir du Mesnil Résidence Le Parc d'Angoulême Lieu-dit Les Vallettes de Clare	Сант-де-л'Абрёвуар-дю-Мёсниль Резиданс-Ле-Парк-д'Ангулем Льё-ди-Ле-Валлет-де-Клар
ICE	Ennisvegur Höfðaströnd Sjávarkambur Þykkvabæjarklaustursvegur	Эннисвегюр-Хёвдастрёнд Сьяуваркамбюр Тикквабайярклёйстюрсвегюр

Расстановка ударений и определение размера стиха

Постановка задачи

Автоматическая расстановка ударений и определение размера стиха в проекте «Стихолюб» *(2009)*

Решение:

- Автоматическая расстановка ударений, фонетическая транскрипция
- Определение схемы стиха:
  - размер (6 силлабо-тонических и 3 тонических размера)
  - клаузула
- Анализ строфы в целом на основе внутренней симметрии, т.е. выбор схемы с наименьшей энтропией

Постановка задачи

Автоматическая расстановка ударений и определение размера стиха в проекте «Стихолюб» *(2009)*

Решение:

- Автоматическая расстановка ударений, фонетическая транскрипция
- Определение схемы стиха:
  - размер (6 силлабо-тонических и 3 тонических размера)
  - клаузула
- Анализ строфы в целом на основе внутренней симметрии, т.е. выбор схемы с наименьшей энтропией

Определение размера стиха

ВАРИАНТ 1 (AAbb)

Швéд, рýсский — кóлет, рýбит, рéжет.		Я4ж
Бой барабáнный, клúки, скрéжет,		Я4ж
Гром пýшек, тóпот, ржáнье, стóн,		Я4м
И смéрть, и áд со всéх сторóн.		Я4м

ВАРИАНТ 2 (AAbb)

Швéд, рýсский — кóлет, рýбит, рéжет.	(0)−0−1−1−1−(1)	Дл5
Бой барабáнный, клúки, скрéжет,		Я4ж
Гром пýшек, тóпот, ржáнье, стóн,		Я4м
И смéрть, и áд со всéх сторóн.		Я4м

Определение размера стиха: новые применения

- Найдено 60 тыс. ошибок в поэтическом корпусе НКРЯ
- Автопоэт – составление стихов из поисковых запросов, твитов, заголовков новостей и т.п.

Пример творчества Автопоэта

курить бамбук что это значит
зачем коту усы и хвост
егэ по химии задачи
тамбов мужчина средний рост

куда пойти в москве в июне
иван тургенев накануне
чикаго платье с бахромой
кальян купить курить домой

все сериалы межсезонья
сломался кухонный комбайн
малифисента фильм онлайн
значенье имени хавронья

редиска перец лук морковь
зачем топтать мою любовь

http://autopoet.yandex.ru

Mystem 3.0

Программа MyStem: новая версия

- **MyStem** — свободно распространяемый морфологический анализатор для русского языка
- http://api.yandex.ru/mystem/
- Первая версия программы была написана Ильёй Сегаловичем в 1997 году
- Мы представляем версию MyStem 3.0
- Её основные отличия:
  - ранжирование разборов,
  - поддержка фикслистов,
  - разные форматы вывода.

Ранжирование разборов

Задача №1 — ранжировать разборы без учета контекста:

есть
есть = V, несов, пе
есть = INTJ
быть = V, нп

Ранжирование разборов

Задача №1 — ранжировать разборы без учета контекста:

есть
есть = V, несов, пе
есть = INTJ
быть = V, нп

айпад
айпада ?= S, муж, неод = вин, ед | им, ед
айпад ?= ADV =
айпад ?= S, муж, од = вин, мн | род, мн
айпада ?= S, жен, неод = род, мн

Ранжирование разборов

Должно получиться примерно так:

есть

2. есть = V, несов, пе = инф
3. есть = INTJ =
1. быть = V, нп = (...)

есть

айпада ?= S, муж, од = (вин, мн | род, мн)

айпад ?= ADV=

1. айпад ?= S, муж, неод = (вин, ед | им, ед)

айпада ?= S, жен, неод = род, мн

Частоты для ранжирования

- Берем из корпуса со снятой омонимией (НКРЯ)
- Учитываем частоты слов из веба

Однако корпус не полный, и запоминать частоты для каждой словоформы расточительно.

Поэтому мы «факторизуем» эти частоты, настраивая их отдельно для
– окончаний каждой схемы,
– основ каждой схемы,
– самих морфологических схем.

Более формально

Пусть зафиксирована схема морфологического разбора ***s***, и в слове ***w*** выделены основа ***stem*** и окончание ***flex***.

Считаем, что события «*встретить основу слова*» и «*встретить окончание слова*» в рамках этой схемы **независимы**.

$$P(scheme \mid word) =$$

$$= \frac{P(word \mid scheme)P(scheme)}{P(word)} =$$

$$= \frac{P(stem \mid scheme)P(flex \mid scheme)P(scheme)}{P(word)}.$$

Оценка качества

Тестировали на подкорпусе НКРЯ со снятой омонимией, выбирая самый вероятную лемму из предложенных.

Стратегии:
– первая лемма: **90%**;
– первая по алфавиту лемма: **89%**;
– самая вероятная лемма в новой модели: **95,5%**!

Снятие омонимии

Опция `-d` включает переранжирование разборов с учётом контекста.

Модель обучена с помощью технологии MatrixNet.
В качестве факторов используются в том числе «нормализующие подстановки» из модели Ю. Зеленкова, И. Сегаловича и В. Титова (Диалог-2005).

Точность первого разбора (по тексту леммы) составляет **97,8%**.

Дополнительные опции

- Опция `--generate-all` строит все гипотезы для неизвестных слов, а не только те, где совпадение с образцом максимально
- Опция `--filter-gram` разрешает строить только разборы с указанными грамматическими тегами

`mystem --filter-gram V`

батарея { батареть ?= V, ipf, intr = Inpraes, ger }

Спасибо за внимание!