

ОСНОВЫ ОБРАБОТКИ ТЕКСТОВ

Лекция #6: Базовые задачи обработки текстов

15 октября 2025

Задача NERC

- На входе: текст, разбитый на предложения и токены
- На выходе: множество сущностей (подстрока, текст)

Александр Пушкин родился в Москве, столице России

PER LOC LOC

Microsoft — один из крупнейших производителей ПО в мире

ORG

Задача NERC

- На входе: текст, разбитый на предложения и токены
- На выходе: множество сущностей (подстрока, текст)

Александр Пушкин родился в Москве, столице России



Александр Пушкин родился в Москве, столице России

PER

LOC

LOC

Microsoft — один из крупнейших производителей ПО в мире



Microsoft — один из крупнейших производителей ПО в мире

ORG

Сегментация текста

- На входе: текст (последовательность символов)
- На выходе: текст, разбитый на предложения и токены

Александр Пушкин родился в Москве, столице России



Александр Пушкин родился в Москве, столице России

Microsoft — один из крупнейших производителей ПО в мире



Microsoft — один из крупнейших производителей ПО в мире

Сегментация текста

- На входе: текст (последовательность символов)
- На выходе: текст, разбитый на предложения и токены
 - Предложение – это единица языка, которая представляет собой грамматически организованное соединение слов, обладающее смысловой законченностью.
 - Токен – минимальная лингвистическая единица текста (речи), обладающая смыслом.

{слово, пунктуация, число}

Токенизация

- На входе: текст (последовательность символов)
- На выходе: текст, разбитый на предложения и токены

Александр Пушкин родился в Москве, столице России.

- Решение (RegExr, регулярные выражения)

Токенизация

- На входе: текст (последовательность символов)
- На выходе: текст, разбитый на предложения и токены

Александр Пушкин родился в Москве, столице России.

- Решение (RegExp, регулярные выражения)

Разделить входную строку по пробельным символам

Александр Пушкин родился в Москве, столице России.

Токенизация

- На входе: текст (последовательность символов)
- На выходе: текст, разбитый на предложения и токены

Александр Пушкин родился в Москве, столице России.

- Решение (RegExp, регулярные выражения)

Разделить входную строку по пробельным символам

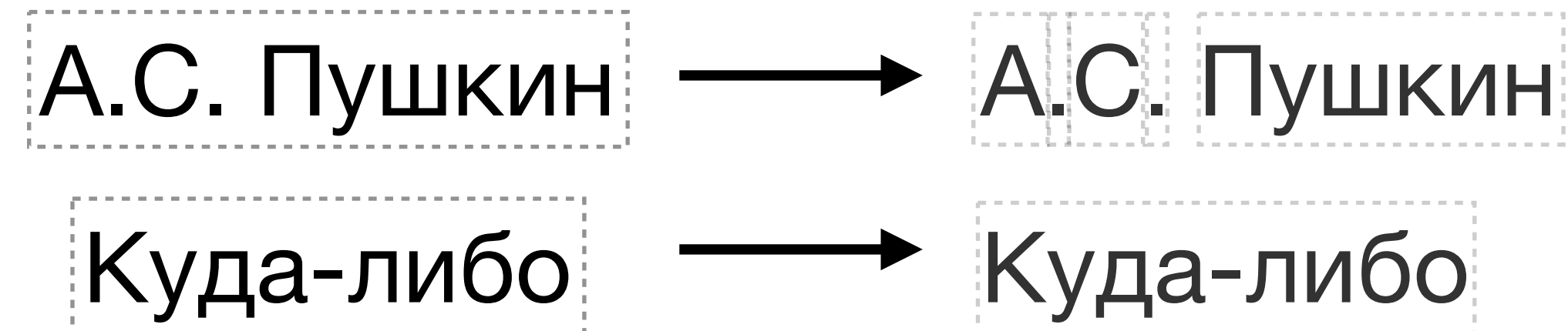
Александр Пушкин родился в Москве, столице России.

Отделить префиксные и постфиксные знаки пунктуации

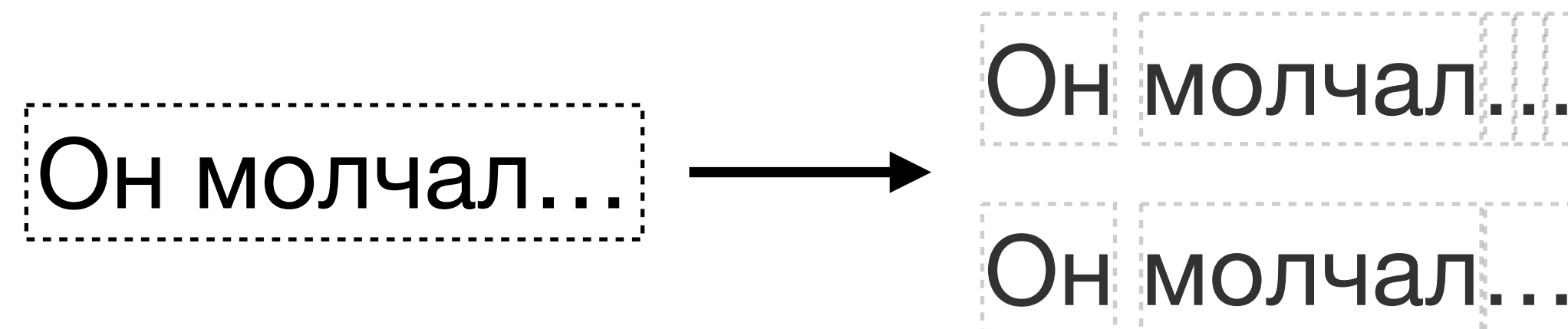
Александр Пушкин родился в Москве, столице России.

Токенизация

- Пунктуация бывает внутри слова



- Может быть несколько символов пунктуации подряд



Токенизация

- Будем считать токенами подстроки, содержащие только буквы и цифры (alphanumeric)
- Остальные подстроки попробуем разделить по не alphanumeric символам, используя бинарный классификатор

А.С.

2.51

Куда-либо

Признаки:

- Длины токенов-кандидатов
- Символы (n-gram)
- Словарь
- ...

Классификаторы:

- SVM
- Лог. регрессия
- Нейронные сети
- ...

Токенизация

- Некоторые токены должны состоять из нескольких слов

- Русский

Как только началось → Как только началось

- Испанский

¿Cómo te llamas? → ¿Cómo te llamas?

Токенизация

- Некоторые слова склеиваются

- Английский

isn't



isn't (is not)

- Испанский

Dímelo



Dímelo (Di me lo)

del



de el

Токенизация

- Некоторые слова склеиваются

- Английский

isn't



isn't

(is not)

- Испанский

Dímelo



Dímelo

(Di me lo)

del



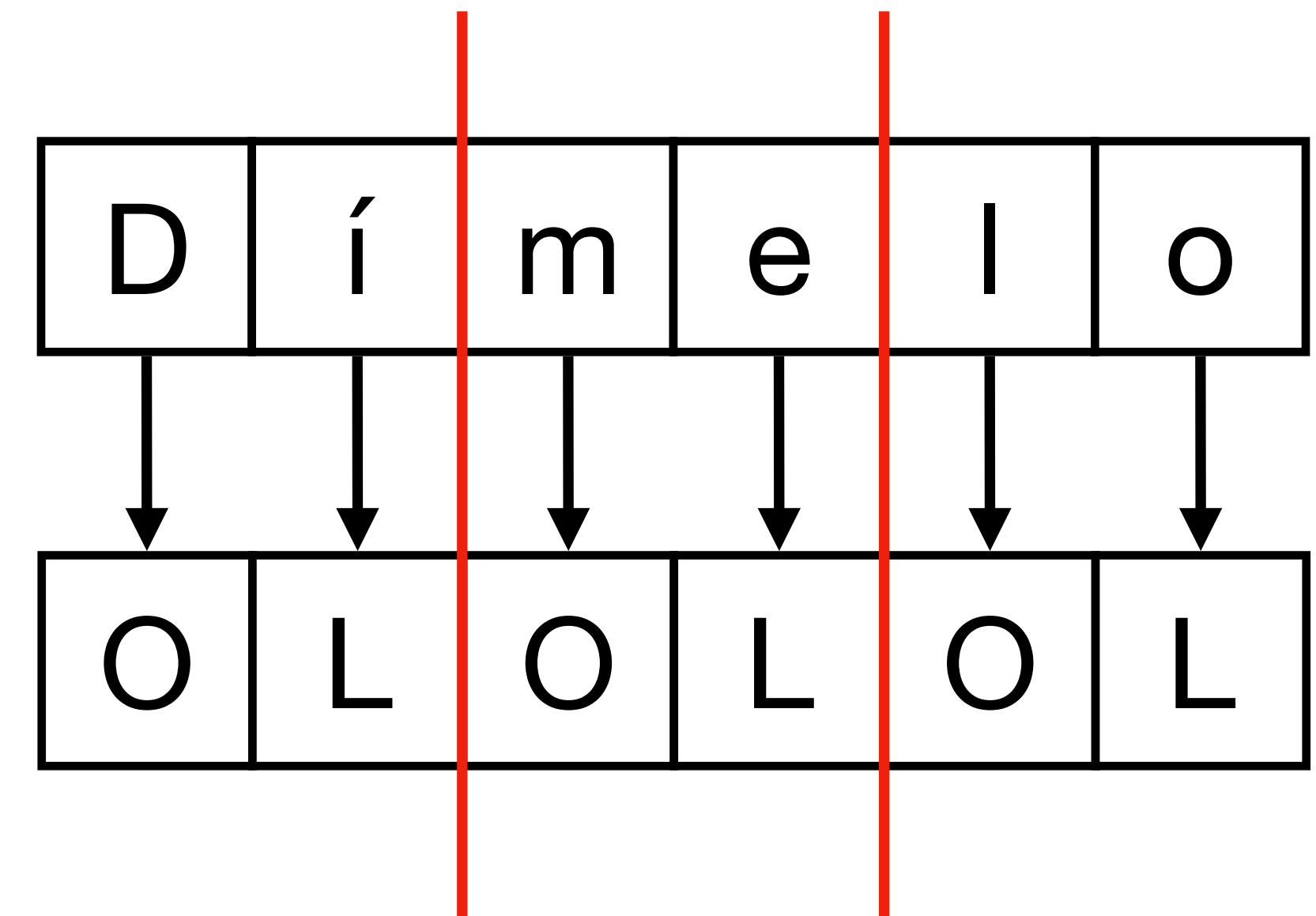
de el

- Немецкий

Rechtsschutzversicherungsgesellschaften

Токенизация

- Задача разметки последовательности. Каждый символ необходимо отнести к одному из двух классов:
 - Последний символ в токене (L)
 - Не последний символ в токене (O)
- Методы:
 - RNN (Рекуррентные нейронные сети)
 - BiRNN
 - ...



- После разделения на символы замена токенов по словарю

Dímelo → (Di me lo)

Определение границ предложений

На входе:

- текст (последовательность токенов)
- текст (последовательность символов)

На выходе: текст, разбитый на предложения

Идея:

- В естественных языках предложения отделяются друг от друга знаками препинания (. ! ?) (。 в китайском) (: в армянском)

Решение:

- Классифицировать каждый токен (. ! ?) является ли он концом предложения

Сегментация текста

На входе: текст (последовательность символов)

На выходе: текст, разбитый на предложения и токены

Идея

Одновременно находить границы токенов и предложений

Решение

- Разметка последовательности. Каждый символ классифицируется на один из **трех** классов:
 - Последний символ предложения
 - Последний символ токена
 - Обычный символ

Оценка качества

- Gold

Александр Пушкин родился в Москве, столице России.

- Predicted

Александр Пушкин родился в Москве, столице России.

Оценка качества

- Gold

Александр Пушкин родился в Москве, столице России.

- Predicted

Александр Пушкин родился в Москве, столице России.

- Метрики

$$P = \frac{|\text{correct}|}{|\text{predicted}|}$$

$$R = \frac{|\text{correct}|}{|\text{gold}|}$$

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}$$

$$P = \frac{5}{7}$$

$$R = \frac{5}{9}$$

$$F_1 = \frac{10}{16}$$

Определение языка

- На входе: текст (последовательность символов)
- На выходе: метка языка

Александр Сергеевич Пушкин родился в Москве
26 мая 1799 года.

Александр Сергей улы Пушкин 1799 елның 26
маенда Мәскәү шәһәрәндә туа.

Аляксандр Сяргеевіч Пушкін нарадзіўся ў
Маскве 26 мая 1799 года.

Alexander Sergej Pushkin fæddist í Moskvu 26 Maí
1799.

ალექსანდრე სერგეევჲი პუშკინი დაიბადა
მოსკოვში 1799 წლის 26 მაისს.

Определение языка

- На входе: текст (последовательность символов)
- На выходе: метка языка

Александр Сергеевич Пушкин родился в Москве
26 мая 1799 года.

русский

Александр Сергей улы Пушкин 1799 елның 26
маенда Мәскәү шәһәрәндә туа.

Аляксандр Сяргеевіч Пушкін нарадзіўся ў
Маскве 26 мая 1799 года.

Alexander Sergej Pushkin fæddist í Moskvu 26 Maí
1799.

ალექსანდრე სერგეევჲი პუშკინი დაიბადა
მოსკოვში 1799 წლის 26 მაისს.

Определение языка

- На входе: текст (последовательность символов)
- На выходе: метка языка

Александр Сергеевич Пушкин родился в Москве
26 мая 1799 года.

русский

Александр Сергей улы Пушкин 1799 елның 26
маенда Мәскәү шәһәрәндә туа.

татарский

Аляксандр Сяргеевіч Пушкін нарадзіўся ў
Маскве 26 мая 1799 года.

Alexander Sergej Pushkin fæddist í Moskvu 26 Maí
1799.

ალექსანდრე სერგეევჲი პუშკინი დაიბადა
მოსკოვში 1799 წლის 26 მაისს.

Определение языка

- На входе: текст (последовательность символов)
- На выходе: метка языка

Александр Сергеевич Пушкин родился в Москве
26 мая 1799 года.

русский

Александр Сергей улы Пушкин 1799 елның 26
маенда Мәскәү шәһәрәндә туа.

татарский

Аляксандр Сяргеевіч Пушкін нарадзіўся ў
Маскве 26 мая 1799 года.

белорусский

Alexander Sergej Pushkin fæddist í Moskvu 26 Maí
1799.

ალექსანდრე სერგეევჲი პუშკინი დაიბადა
მოსკოვში 1799 წლის 26 მაისს.

Определение языка

- На входе: текст (последовательность символов)
- На выходе: метка языка

Александр Сергеевич Пушкин родился в Москве
26 мая 1799 года.

русский

Александр Сергей улы Пушкин 1799 елның 26
маенда Мәскәү шәһәрәндә туа.

татарский

Аляксандр Сяргеевіч Пушкін нарадзіўся ў
Маскве 26 мая 1799 года.

белорусский

Alexander Sergej Pushkin fæddist í Moskvu 26 Maí
1799.

исландский

ალექსანდრე სერგეევჲი პუშკინი დაიბადა
მოსკოვში 1799 წლის 26 მაისს.

Определение языка

- На входе: текст (последовательность символов)
- На выходе: метка языка

Александр Сергеевич Пушкин родился в Москве
26 мая 1799 года.

русский

Александр Сергей улы Пушкин 1799 елның 26
маенда Мәскәү шәһәрәндә туа.

татарский

Аляксандр Сяргеевіч Пушкін нарадзіўся ў
Маскве 26 мая 1799 года.

белорусский

Alexander Sergej Pushkin fæddist í Moskvu 26 Maí
1799.

исландский

ალექსანდრე სერგეევინი პუშკინი დაიბადა
მოსკოვში 1799 წლის 26 მაისს.

грузинский

Профили языка*

Для каждого языка извлекается его профиль

- Профиль – список (top-K) символьных n-gram отсортированный по невозрастанию частоты

Происхождение Александра Сергеевича Пушкина идёт от разветвлённого нетитулованного дворянского рода Пушкиных, восходившего по генеалогической легенде к «мужу честну» Ратше. Пушкин неоднократно писал о своей родословной в стихах и прозе; он видел в своих предках образец истинной «аристократии», древнего рода, честно служившего отечеству, но не снискавшего благосклонности правителей и «гонимого». Не раз он обращался (в том числе в художественной форме) и к образу своего прадеда по матери — африканца Абрама Петровича Ганнибала, ставшего слугой и воспитанником Петра I, а потом военным инженером и генералом.



' '	81
'o'	66
'e'	49
'н'	42
'a'	41
'и'	37
'р'	28
...	...



' '
'o'
'e'
'н'
'a'

* [W. B. Cavnar, J. M. Trenkle. "N-gram-based text categorization." \(1994\)](#)

Профили языка

Для каждого языка извлекается его профиль

- Профиль – список (top-K) символьных n-gram отсортированный по невозрастанию частоты

Үсмер чагының 6 елын Пушкин 1811 елның 19 октябрендә дәрәжәле кешеләр өчен ачылган лицейда үткәрә. Лицейда укыгында усмернең таланты ачыла һәм башкалар тарафыннан югары бәяләнә. Лицейда үткәрелгән еллар, дуслары шагыйрь күңелендә мәңгегә саклана һәм ижатында чагылыш таба. Александр Сергеевичның яшьлегә шушында үтә, биредә шагыйрь Пушкин туа: 130лап шигырь ижәт ителә, «Руслан һәм Людмила» поэмасы языла башлай, танылган журналларда беренче язмалары басыла. Лицей хәзерге урта яки югары уку йортлары дәрәжәсендә белем бирергә тиеш була.



' '	73
'a'	53
'e'	37
'л'	33
'р'	30
'н'	28
'ы'	28
...	...



' '
'a'
'e'
'л'
'р'

Профили языка

Для классифицируемого текста извлекается профиль

Весной 1820 года Пушкина вызвали к военному генерал-губернатору Петербурга графу М. А. Милорадовичу для объяснения по поводу содержания его стихотворений (в том числе эпиграмм на Аракчеева, архимандрита Фотия и самого Александра I), несовместимых со статусом государственного чиновника.



' '	34
'o'	26
'a'	22
'н'	18
'e'	18
'и'	17
'р'	15
...	...



' '
'o'
'a'
'н'
'e'

Профили языка

Профиль классифицируемого текста сравнивается с профилями языков (out-of-place), выбирается ближайший язык

Профиль языка	Профиль документа	Расстояние
' '	' '	0
'o'	'o'	0
'e'	a'	2
'н'	'н'	0
'a'	'e'	2
		4

русский

Профиль языка	Профиль документа	Расстояние
' '	' '	0
'a'	'o'	5
'e'	a'	1
'л'	'н'	5
'р'	'e'	2
		13

татарский

Naïve Bayes Classifier

Теория

- Теорема Байеса

$$p(c | d) = \frac{p(d | c) \cdot p(c)}{p(d)}$$

- $p(c | d)$ - вероятность $d \in c$
 - $p(d | c)$ - вероятность встретить d в c
 - $p(c)$ - вероятность встретить документ класса c
 - $p(d)$ - вероятность встретить документ d
- Классификатор

$$\hat{c} = \arg \max_{c \in C} p(c | d) = \arg \max_{c \in C} \frac{p(d | c) \cdot p(c)}{p(d)} = \arg \max_{c \in C} p(d | c) \cdot p(c)$$

Naïve Bayes Classifier

Теория

$$\hat{c} = \arg \max_{c \in C} p(d | c) \cdot p(c)$$

- Представим d как $[f_1, f_2, \dots, f_n]$
(набор признаков)

$$\begin{aligned} p(d | c) &= p(f_1, f_2, \dots, f_n | c) = p(f_1 | c) \cdot p(f_2, \dots, f_n | c, f_1) = \\ &= p(f_1 | c) \cdot p(f_2 | c, f_1) \cdot \dots \cdot p(f_n | c, f_1, \dots, f_{n-1}) \end{aligned}$$

Naïve Bayes Classifier

Теория

$$\hat{c} = \arg \max_{c \in C} p(d | c) \cdot p(c)$$

- Представим d как $[f_1, f_2, \dots, f_n]$
(набор признаков)

$$\begin{aligned} p(d | c) &= p(f_1, f_2, \dots, f_n | c) = p(f_1 | c) \cdot p(f_2, \dots, f_n | c, f_1) = \\ &= p(f_1 | c) \cdot p(f_2 | c, f_1) \cdot \dots \cdot p(f_n | c, f_1, \dots, f_{n-1}) \end{aligned}$$

- Предположим, что f_i независимы при условии c

$$p(f_i | c, f_j) = p(f_i | c), i \neq j$$

Naïve Bayes Classifier

Теория

$$\hat{c} = \arg \max_{c \in C} p(d | c) \cdot p(c)$$

- Представим d как $[f_1, f_2, \dots, f_n]$
(набор признаков)

$$p(d | c) = p(f_1 | c) \cdot p(f_2 | c) \cdot \dots \cdot p(f_n | c) = \prod_{i=1}^n p(f_i | c)$$

- Предположим, что f_i независимы при условии c

$$p(f_i | c, f_j) = p(f_i | c), i \neq j$$

Naïve Bayes Classifier

Теория

$$\hat{c} = \arg \max_{c \in C} p(d | c) \cdot p(c)$$

- Представим d как $[f_1, f_2, \dots, f_n]$
(набор признаков)

$$\hat{c} = \arg \max_{c \in C} \prod_{i=1}^n p(f_i | c) \cdot p(c)$$

$$\hat{c} = \arg \max_{c \in C} \left[\sum_{i=1}^n \log p(f_i | c) + \log p(c) \right]$$

Определение языка текста

Naïve Bayes Classifier

$$\hat{c} = \arg \max_{c \in C} \left[\sum_{i=1}^n \log p(f_i | c) + \log p(c) \right]$$

- Признаки — символьные n-gram'ы
- Оценка параметров:

$$p(c) = \frac{|D_c|}{|D|}; \text{ или } p(c) = \frac{1}{|C|};$$

$$p(f_i | c) = \frac{|\{f_i \in D_c\}|}{\sum_j |\{f_j \in D_c\}|}$$

Naïve Bayes Classifier

Сглаживание

- Проблема: некоторые n-gram'ы могут не встретиться в обучающем корпусе

$$p(f_i | c) = \frac{|\{f_i \in D_c\}|}{\sum_j |\{f_j \in D_c\}|} = \frac{0}{\sum_j |\{f_j \in D_c\}|} = 0$$

- Решение: аддитивное сглаживание

$$p(f_i | c) = \frac{|\{f_i \in D_c\}| + \varepsilon}{\sum_j (|\{f_j \in D_c\}| + \varepsilon)} \neq 0$$

Naïve Bayes Classifier

Обучение

Происхождение Александра Сергеевича Пушкина идёт от разветвлённого нетитулованного дворянского рода Пушкиных, восходившего по генеалогической легенде к «мужу честну»

русский

Ратш
стиха
«арис
но не
Не ра
обра
Петр
а пот

Үсмер чагының 6 елын Пушкин 1811 елның 19 октябрэндә дәрәжәле кешеләр өчен ачылган лицейда үткәрә. Лицейда укыгында усмернең таланты ачыла һәм башкалар тарафыннан югары бәяләнә. Лицейда үткәрелгән еллар, дуслары шагыйрь күңелендә мәңгегә саклана һәм ижатында чагылыш таба. Александр Сергеевичның яшьлеге шушында үтә, биредә шагыйрь Пушкин туа: 130лап шигырь ижат ителә, «Руслан һәм Людмила» поэмасы языла башлый, танылган журналларда беренче язмалары басыла. Лицей хәзерге урта яки югары уку йортлары дәрәжәсендә белем бирергә тиеш була.

татарский

$$p(\text{русский}) = \frac{1}{2}$$

$$p(\text{татарский}) = \frac{1}{2}$$

Naïve Bayes Classifier

Обучение

русский

Происхождение Александра Сергеевича Пушкина идёт от разветвлённого нетитулованного дворянского рода Пушкиных, восходившего по генеалогической легенде к «мужу честну» Ратше. Пушкин неоднократно писал о своей родословной в стихах и прозе; он видел в своих предках образец истинной «аристократии», древнего рода, честно служившего отечеству, но не снискавшего благосклонности правителей и «гонимого». Не раз он обращался (в том числе в художественной форме) и к образу своего прадеда по матери — африканца Абрама Петровича Ганнибала, ставшего слугой и воспитанником Петра I, а потом военным инженером и генералом.

$$p(\text{П} \mid \text{русский}) = \frac{6 + 1}{610 + 46} = 0.01067$$

$$p(\text{и} \mid \text{русский}) = \frac{37 + 1}{610 + 46} = 0.05793$$

$$p(\text{р} \mid \text{русский}) = \frac{28 + 1}{610 + 46} = 0.04421$$

$$p(\text{с} \mid \text{русский}) = \frac{27 + 1}{610 + 46} = 0.04268$$

...

$$p(\text{о} \mid \text{русский}) = \frac{66 + 1}{610 + 46} = 0.10213$$

$$p(\text{ж} \mid \text{русский}) = \frac{5 + 1}{610 + 46} = 0.00914$$

Naïve Bayes Classifier

Обучение

татарский

Үсмер чагының 6 елын Пушкин 1811 елның 19 октябрэндә дәрәжәле кешеләр өчен ачылган лицейда үткәрә. Лицейда укыгында усмернең таланты ачыла һәм башкалар тарафыннан югары бәяләнә. Лицейда үткәрелгән еллар, дуслары шагыйрь күңелендә мәңгегә саклана һәм ижатында чагылыш таба. Александр Сергеевичның яшьлеге шушында үтә, биредә шагыйрь Пушкин туа: 130лап шигырь ижәт ителә, «Руслан һәм Людмила» поэмасы языла башлый, танылган журналларда беренче язмалары басыла. Лицей хәзерге урта яки югары уку йортлары дәрәжәсендә белем бирергә тиеш була.

$$p(\text{П} \mid \text{русский}) = \frac{2 + 1}{538 + 54} = 0.00507$$

$$p(\text{и} \mid \text{русский}) = \frac{16 + 1}{538 + 54} = 0.02872$$

$$p(\text{р} \mid \text{русский}) = \frac{30 + 1}{538 + 54} = 0.05236$$

$$p(\text{с} \mid \text{русский}) = \frac{9 + 1}{538 + 54} = 0.01689$$

...

$$p(\text{о} \mid \text{русский}) = \frac{3 + 1}{538 + 54} = 0.00675$$

$$p(\text{ж} \mid \text{русский}) = \frac{1 + 1}{538 + 54} = 0.00337$$

Naïve Bayes Classifier

Предсказание

Весной 1820 года Пушкина вызвали к военному генерал-губернатору Петербурга графу М. А. Милорадовичу для объяснения по поводу содержания его стихотворений (в том числе эпиграмм на Аракчеева, архимандрита Фотия и самого Александра I), несовместимых со статусом государственного чиновника.

$$\hat{c} = \arg \max_{c \in C} \left[\sum_{i=1}^n \log p(f_i | c) + \log p(c) \right]$$

$$\hat{c}_{\text{русский}} = \begin{matrix} \log p(c) & \log p(B|c) & \log p(e|c) & \log p(c|c) & \log p(.|c) \\ -0.69315 & -6.48768 & -2.57566 & -3.15547 & -5.10139 \end{matrix}$$

$$\hat{c}_{\text{русский}} = -953.05578$$

$$\hat{c}_{\text{татарский}} = \begin{matrix} \log p(c) & \log p(B|c) & \log p(e|c) & \log p(c|c) & \log p(.|c) \\ -0.69315 & -6.40026 & -2.76267 & -4.09767 & -4.60850 \end{matrix}$$

$$\hat{c}_{\text{татарский}} = -1066.18426$$

Naïve Bayes Classifier

Предсказание

- Назад к вероятностям

$$p(c_i | d) = \frac{\exp \hat{c}_i}{\sum_j \exp \hat{c}_j} = \frac{1}{1 + \sum_{j \neq i} \exp \hat{c}_j - \hat{c}_i}$$

- Наш пример

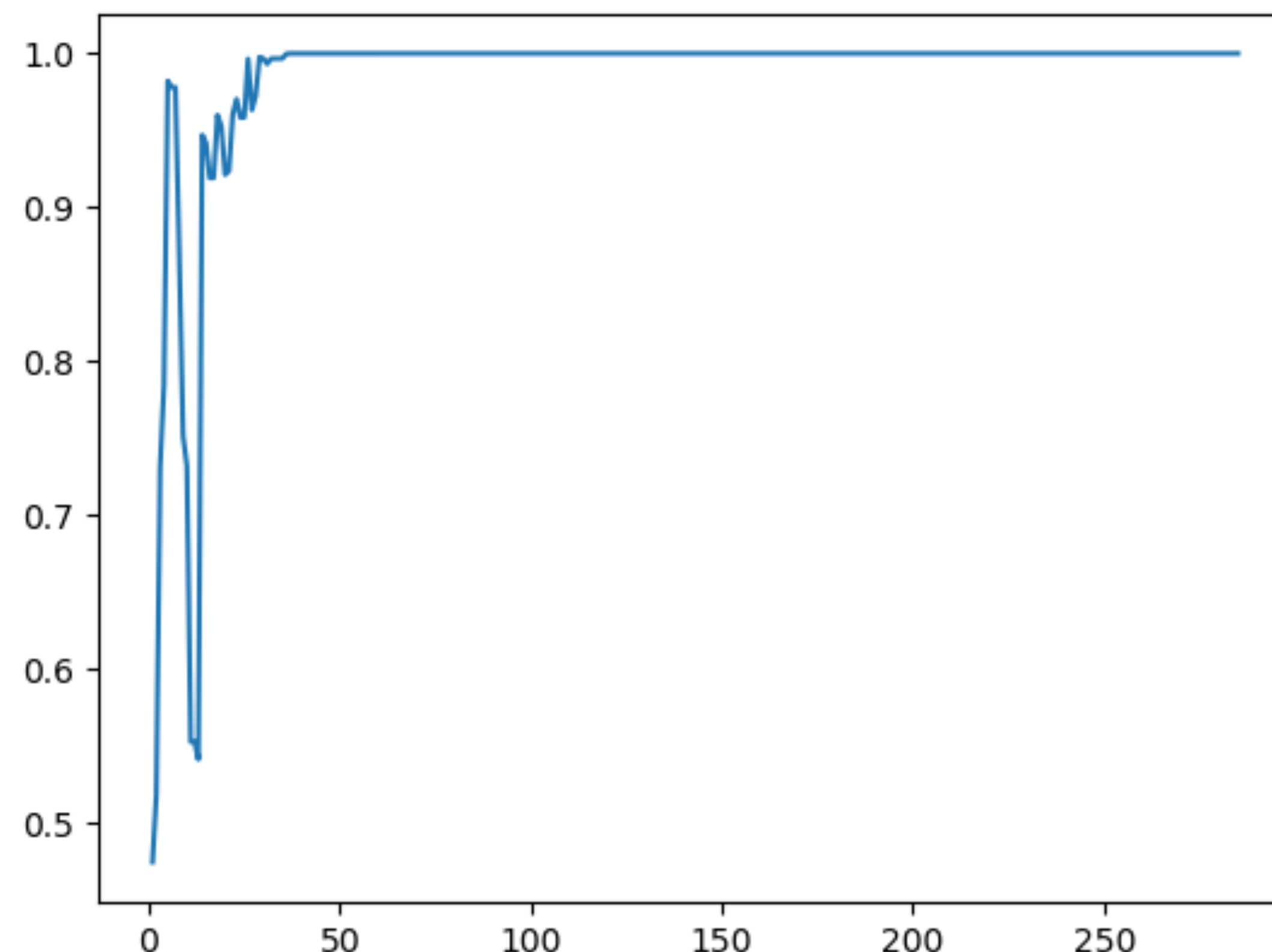
- $\hat{c}_{\text{русский}} = -953.05578$; $\hat{c}_{\text{татарский}} = -1066.1$

- $p(\text{русский} | d) = \frac{1}{1 + e^{-113.12848}} \approx 1$

- Для первых 30 символов

- $\hat{c}_{\text{русский}} = -113.25854$; $\hat{c}_{\text{татарский}} = -118.98256$

- $p(\text{русский} | d) = \frac{1}{1 + e^{-5.72402}} \approx 0.99674$



Весной 1820 года Пушкина вызва

Определение языка

- Задача многоклассовой классификации (multiclass classification)
 - Объекты — тексты
 - Классы — метки языков
- Признаки классификации
 - N-gram по символам
 - Слова (из словаря)
- Классификаторы
 - Multiclass SVM
 - Naïve Bayes
 - Нейронные сети
 - Perceptron
 - Convolutional NN)
 - ...

Оценка качества

Binary classification

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

		Ожидаемые классы	
		спам	не спам
Предсказанные классы	спам	tp	fp
	не спам	fn	tn

Confusion table

Оценка качества

multiclass classification

		Ожидаемые классы		
		русский	татарский	английский
Предсказанные классы	русский	5	2	0
	татарский	3	3	2
	английский	0	1	10

Confusion matrix

Оценка качества

multiclass classification

		Ожидаемые классы		
		русский	татарский	английский
Предсказанные классы	русский	5	2	0
	татарский	3	3	2
	английский	0	1	10

Confusion matrix

Precision = $\frac{5}{7}$

Recall = $\frac{5}{8}$

$F_1 = \frac{2}{3}$

		Ожидаемые классы	
		русский	не русский
Предсказанные классы	русский	5	2
	не русский	3	17

Confusion table

Оценка качества

multiclass classification

		Ожидаемые классы		
		русский	татарский	английский
Предсказанные классы	русский	5	2	0
	татарский	3	3	2
	английский	0	1	10

Confusion matrix

Precision = $\frac{3}{8}$

Recall = $\frac{1}{2}$

$F_1 = \frac{3}{7}$

		Ожидаемые классы	
		татарский	не татарский
Предсказанные классы	татарский	3	5
	не татарский	3	15

Confusion table

Оценка качества

multiclass classification

		Ожидаемые классы		
		русский	татарский	английский
Предсказанные классы	русский	5	2	0
	татарский	3	3	2
	английский	0	1	10

Confusion matrix

Precision = $\frac{10}{11}$

Recall = $\frac{5}{6}$

$F_1 = \frac{20}{23}$

		Ожидаемые классы	
		английский	не английский
Предсказанные классы	английский	10	1
	не английский	2	13

Confusion table

Оценка качества

multiclass classification

Макро усреднение

Precision = $\frac{1}{|C|} \sum \text{Precision}^c$

Recall = $\frac{1}{|C|} \sum \text{Recall}^c$

$F_1 = \frac{1}{|C|} \sum F_1^c$

Precision = $\frac{1231}{1848} \approx 0.6661$

Recall = $\frac{47}{72} \approx 0.6528$

$F_1 = \frac{949}{1449} \approx 0.6549$

		Ожидаемые классы	
		русский	не русский
Предсказанные классы	русский	5	2
	не русский	3	17

		Ожидаемые классы	
		татарский	не татарский
Предсказанные классы	татарский	3	5
	не татарский	3	15

		Ожидаемые классы	
		английский	не английский
Предсказанные классы	английский	10	1
	не английский	2	13

Оценка качества

multiclass classification

Взвешенное макро усреднение

$$\alpha_c = tp_c + fn_c$$

Precision = $\frac{\sum \alpha_c \cdot \text{Precision}^c}{\sum \alpha_c}$

Recall = $\frac{\sum \alpha_c \cdot \text{Recall}^c}{\sum \alpha_c}$

$F_1 = \frac{\sum \alpha_c \cdot F_1^c}{\sum \alpha_c}$

Precision = $\frac{5813}{8008} \approx 0.7259$

Recall = $\frac{9}{13} \approx 0.6923$

$F_1 = \frac{4429}{6279} \approx 0.7054$

		Ожидаемые классы	
		русский	не русский
Предсказанные классы	русский	5	2
	не русский	3	17

		Ожидаемые классы	
		татарский	не татарский
Предсказанные классы	татарский	3	5
	не татарский	3	15

		Ожидаемые классы	
		английский	не английский
Предсказанные классы	английский	10	1
	не английский	2	13

Определение языка

Сложности

- Очень короткий текст

Один два три четыре пять

русский

Адзін два тры чатыры пяць

белорусский

Один два три чотири п'ять

украинский

Един два три четири пет

болгарский

Еден два три четири пет

македонский

Један два три четири пет

сербский

Определение языка

Сложности

- Текст сразу на нескольких языках

I heard “El que quiera entender que entienda” by Mägo de Oz?

испанский

английский

¿Conoces “Born to Touch Your Feelings” de Scorpions?

английский

испанский

I am at Международный аэропорт Пулково in Санкт-Петербург

английский

русский

Определение языка

Сложности

- Текст сразу на нескольких языках

I **heard** “El que quiera entender que entienda” by Mägo de Oz?

английский

¿**Conoces** “Born to Touch Your Feelings” de Scorpions?

испанский

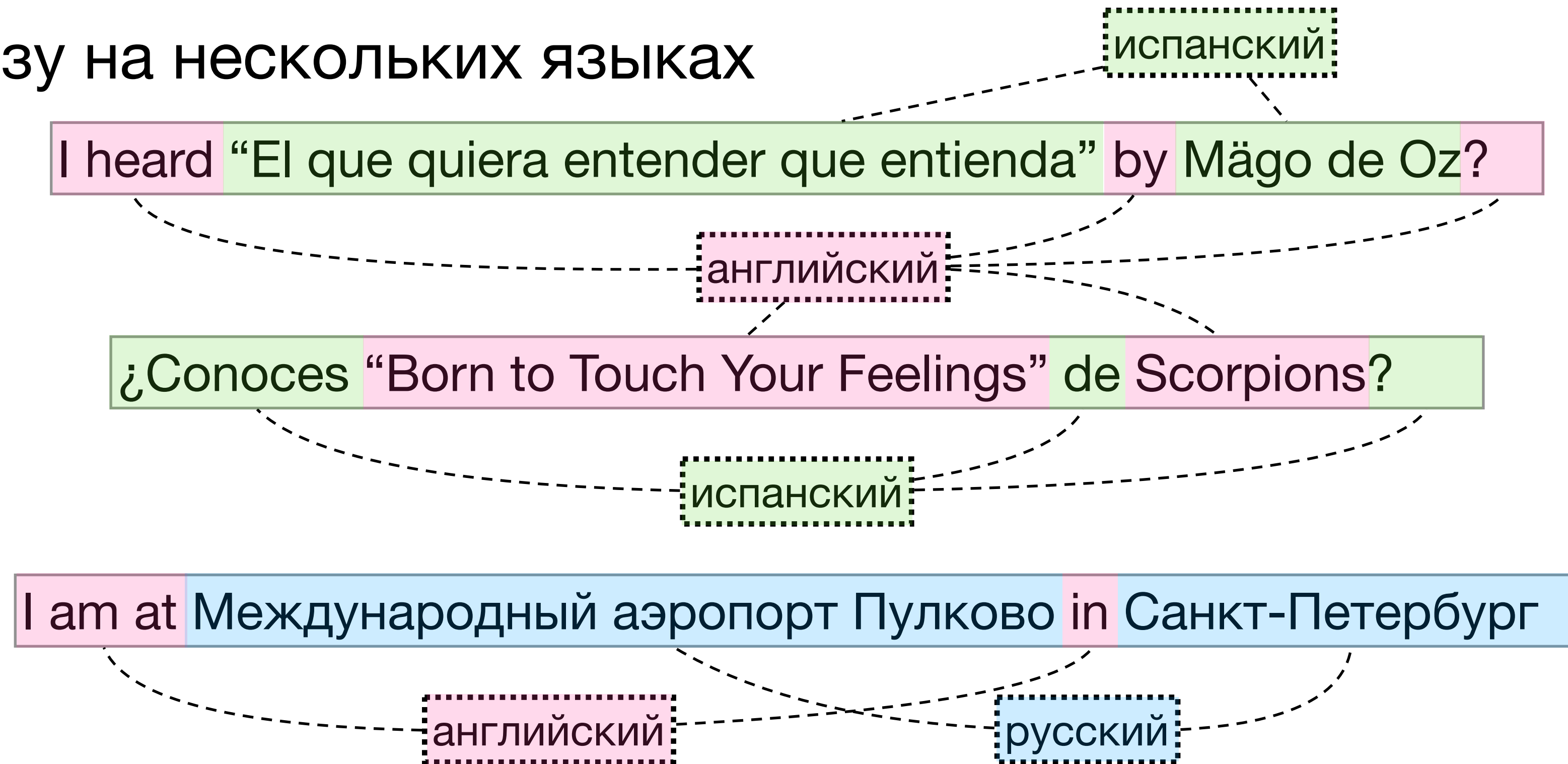
I **am** at Международный аэропорт Пулково in Санкт-Петербург

английский

Определение языка

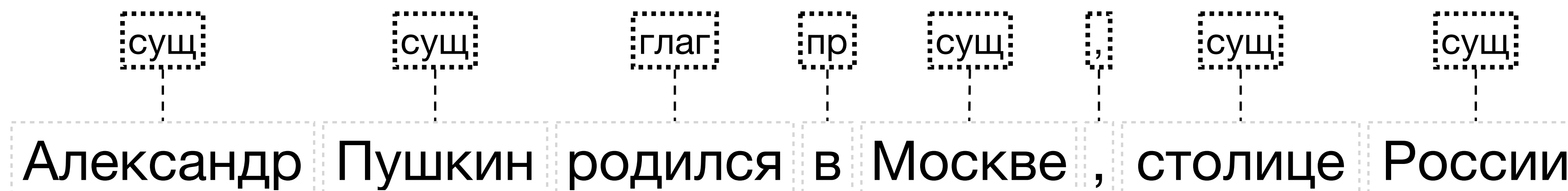
Сложности

- Текст сразу на нескольких языках



Морфологический анализ

- Задача
 - На входе: предложение (последовательность токенов)
 - На выходе: морфологическая метка для каждого слова
- Морфологическая метка – часть речи
 - Существительное
 - Прилагательное
 - Глагол
 - ...



Морфологический анализ

Задача разметки последовательности

- На входе: последовательность токенов
- На выходе: морфологическая метка для каждого слова

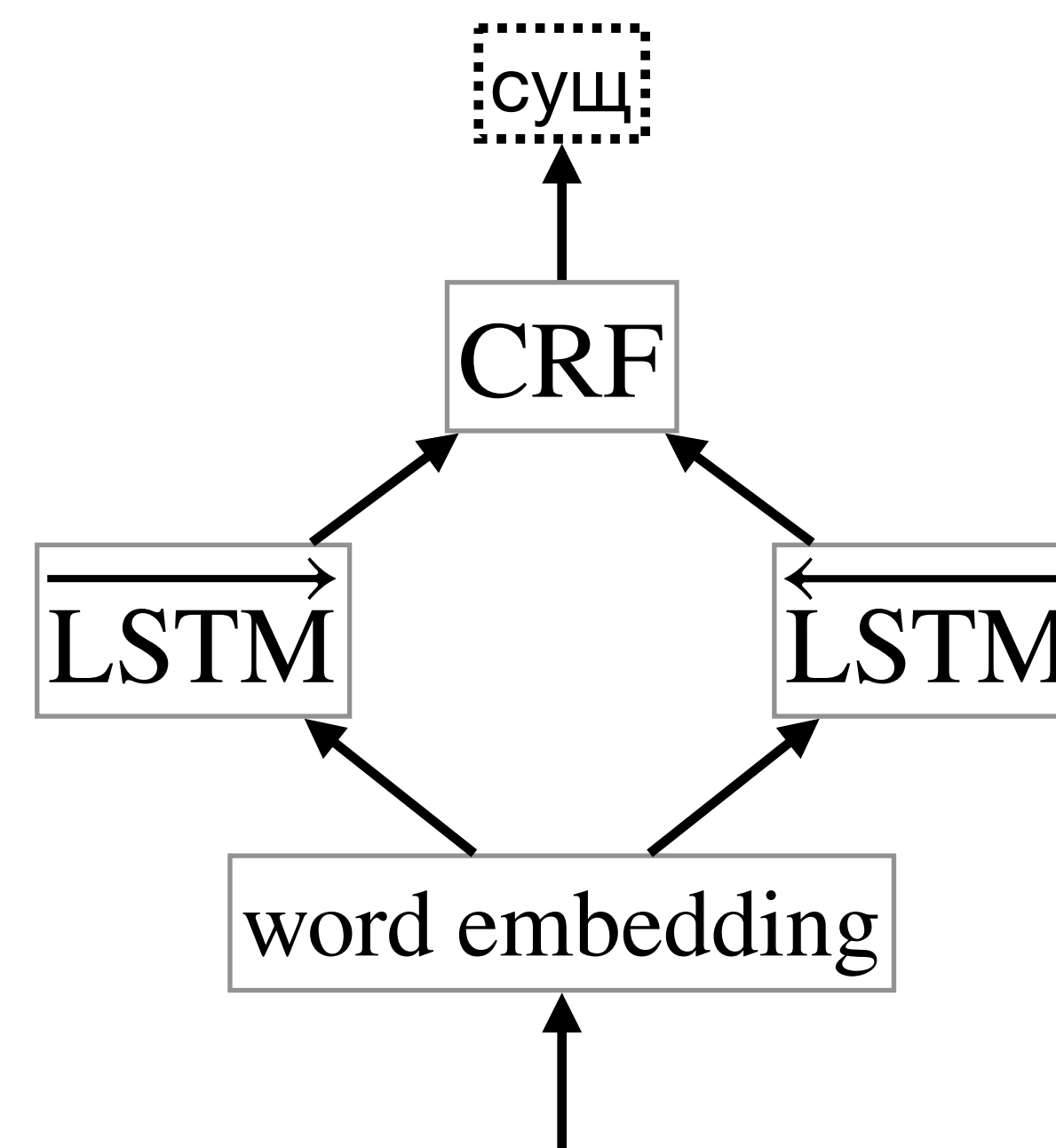
Методы

- CRF
- RNN (biRNN)
- biRNN-CRF

Признаки

- Слова (векторные представления)
- Части слов (префиксы, суффиксы)

Александр Пушкин родился в Москве , столице России



Морфологический анализ

Методы оценки качества

- Метрики многоклассовой классификации
 - Precision, Recall, F_1
- Sentence Accuracy
 - Количество правильно разобранных предложений среди всех предложений

Морфологический анализ

- Задача
 - На входе: предложение (последовательность токенов)
 - На выходе: морфологическая метка для каждого слова
- Морфологическая метка – часть речи + граммемы

Части речи

- Существительное
- Прилагательное
- Глагол
- ...

Род

- Мужской
- Женский
- Средний

Число

- Единственное
- Множественное

Падеж

- Именительный
- Родительный
- Дательный
- ...

...

сущ м.р. ед.ч
Им.п. одуш.

Александр

сущ м.р. ед.ч
Им.п. одуш.

Пушкин

глагол пр.вр.
инд. ед.ч. м.р.

родился

пр

в

сущ ж.р. ед.ч
Пр.п. неод.

Москве

,

сущ ж.р. ед.ч
Пр.п. неод.

столице

сущ ж.р. ед.ч
Р.п. неод.

России

Морфологический анализ

Проблемы

- Много классов (UD SynTagRus: 733)
- Некоторые классы слишком редкие $\Rightarrow$ не получится обучиться

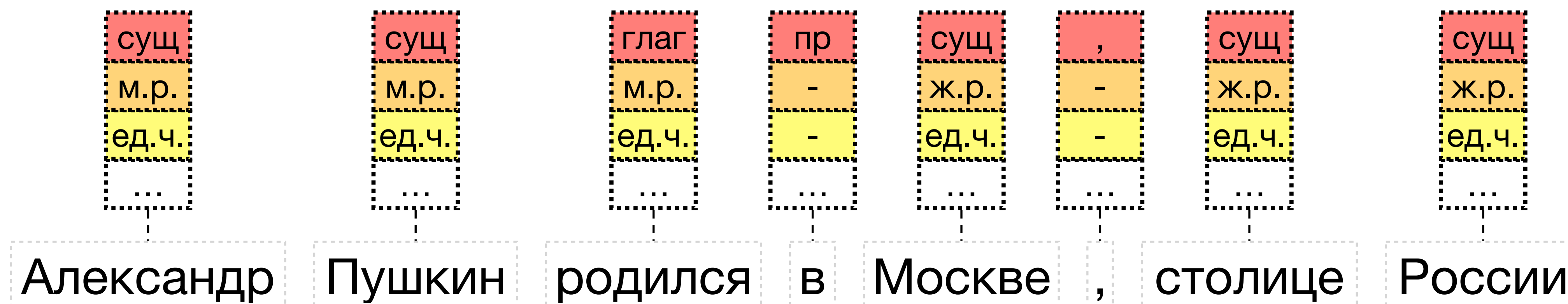
Морфологический анализ

Проблемы

- Много классов (UD SynTagRus: 733)
- Некоторые классы слишком редкие $\Rightarrow$ не получится обучиться

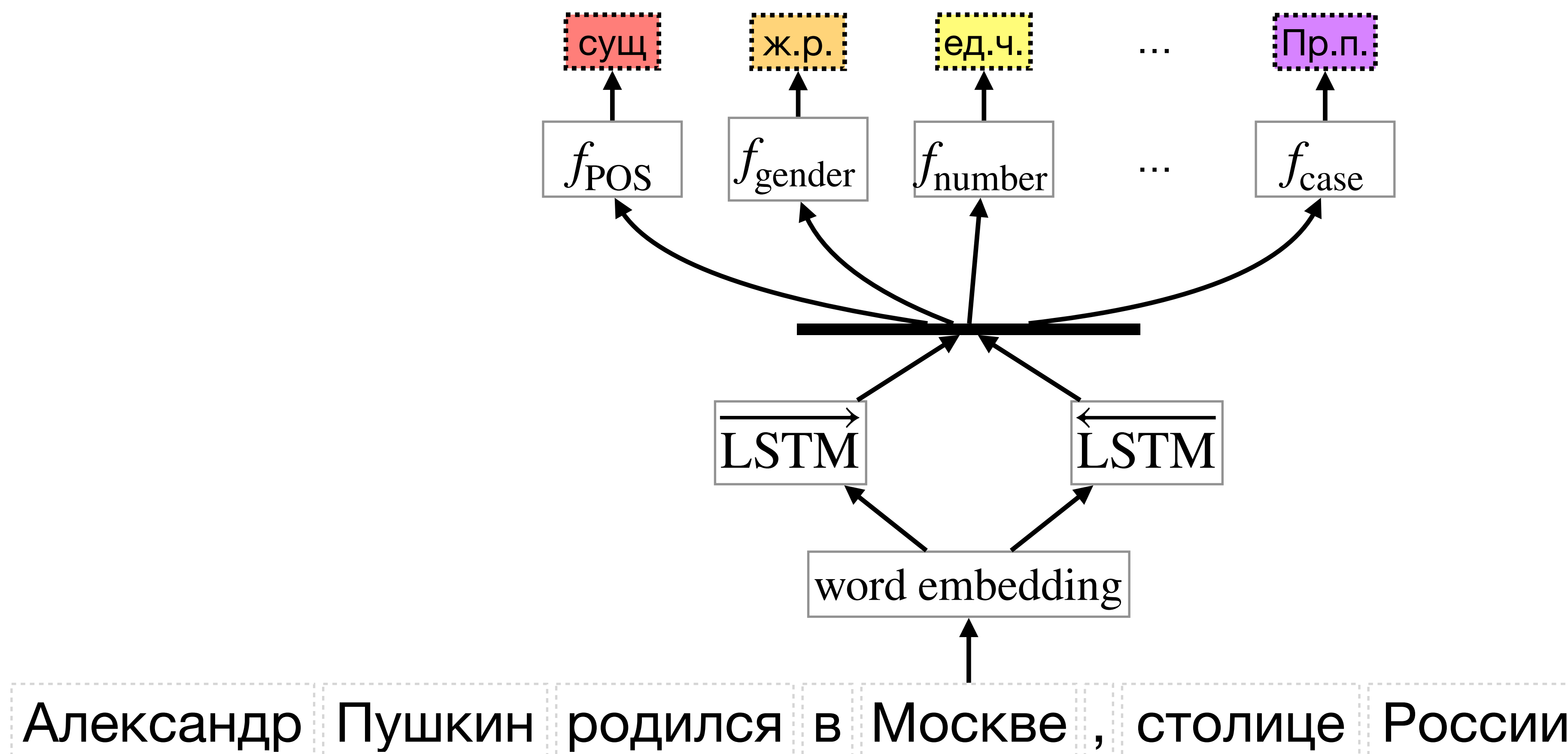
Решение

- Разделить метки по грамматическим категориям
- Назначить каждому токену метку из каждой категории
- Multilabel classification



Морфологический анализ

Multilabel classification



Морфологический анализ

Multilabel classification

При таком подходе может получиться неправильная метка:

- Род, число падеж у предлогов/союзов
- Род у глаголов настоящего времени
- Время у существительных

Решение:

- Собрать множество возможных меток (из корпуса/руками)
- Определять наиболее вероятную метку из возможных

$$p(\text{сущ, м.р, ... , Пр.п} | w) = p(\text{сущ} | w) \cdot p(\text{м.р} | w) \cdot \dots \cdot p(\text{Пр.п} | w)$$

Нормализация (лемматизация) слов

Задача

- На входе: предложение/последовательность слов (токенов)
- На выходе: последовательность лемм – слов (токенов) в нормальной форме

Нормальная форма слова – каноническая форма слова

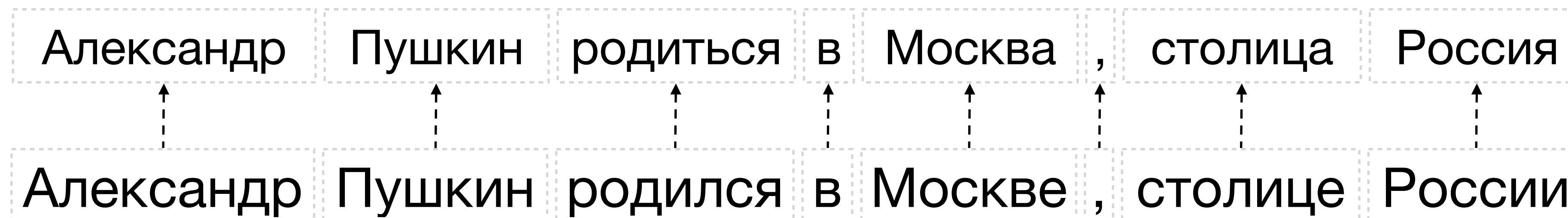
- Для существительных:
 - единственное число
 - именительный падеж
- Для глаголов:
 - инфинитив
- ...

Нормализация (лемматизация) слов

Задача

- На входе: предложение/последовательность слов (токенов)
- На выходе: последовательность лемм – слов (токенов) в нормальной форме

Задача похожа на разметку последовательности



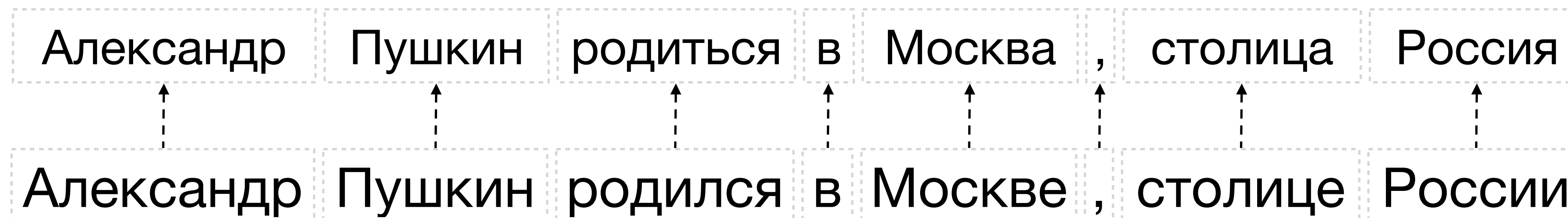
Нормализация (лемматизация) слов

Задача

- На входе: предложение/последовательность слов (токенов)
- На выходе: последовательность лемм – слов (токенов) в нормальной форме

Задача похожа на разметку последовательности, но

- На выходе слова, а не классы
- Слова на выходе не связаны между собой



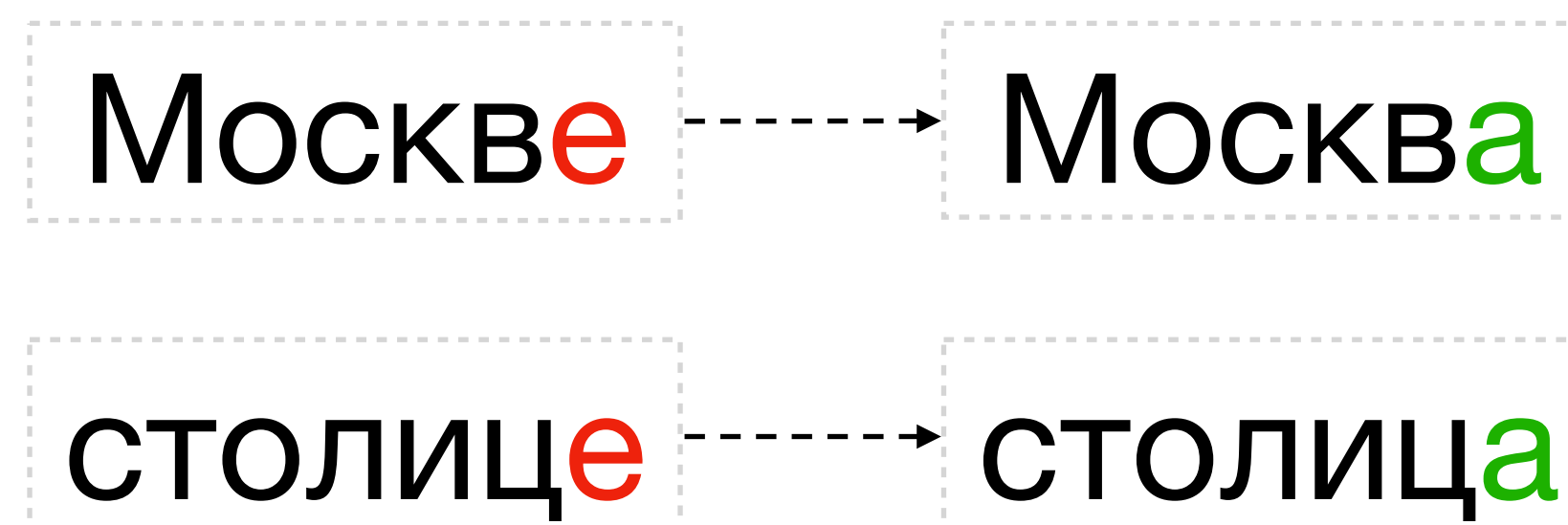
Нормализация (лемматизация) слов

Задача

- На входе: слово (токен)
- На выходе: лемма — слово (токен) в нормальной форме

Процесс построения леммы

- Удалить суффикс словоформы
- Добавить суффикс леммы
- Удалить префикс словоформы
- Добавить префикс леммы



Нормализация (лемматизация) слов

Методы, основанные на правилах

- LemmaGen (2010)

```
IF suffix THEN transformation EXCEPT exceptions
```

```
IF “” THEN “”→”” EXCEPT  
  IF “d” THEN “d”→””  
  ELSE IF “ote” THEN “ote”→”ite”  
  ELSE IF “ing” THEN “ing”→”e” EXCEPT  
    IF “ting” THEN “”→””  
  ELSE IF “ten” THEN “ten”→”e”  
  ELSE IF “s” THEN “s”→””
```

...

Нормализация (лемматизация) слов

Многоклассовая классификация

Классы:

- Суффиксы словоформ, суффиксы лемм, префиксы словоформ, префиксы лемм
- Правила преобразования суффикса, правила преобразования префикса

Признаки:

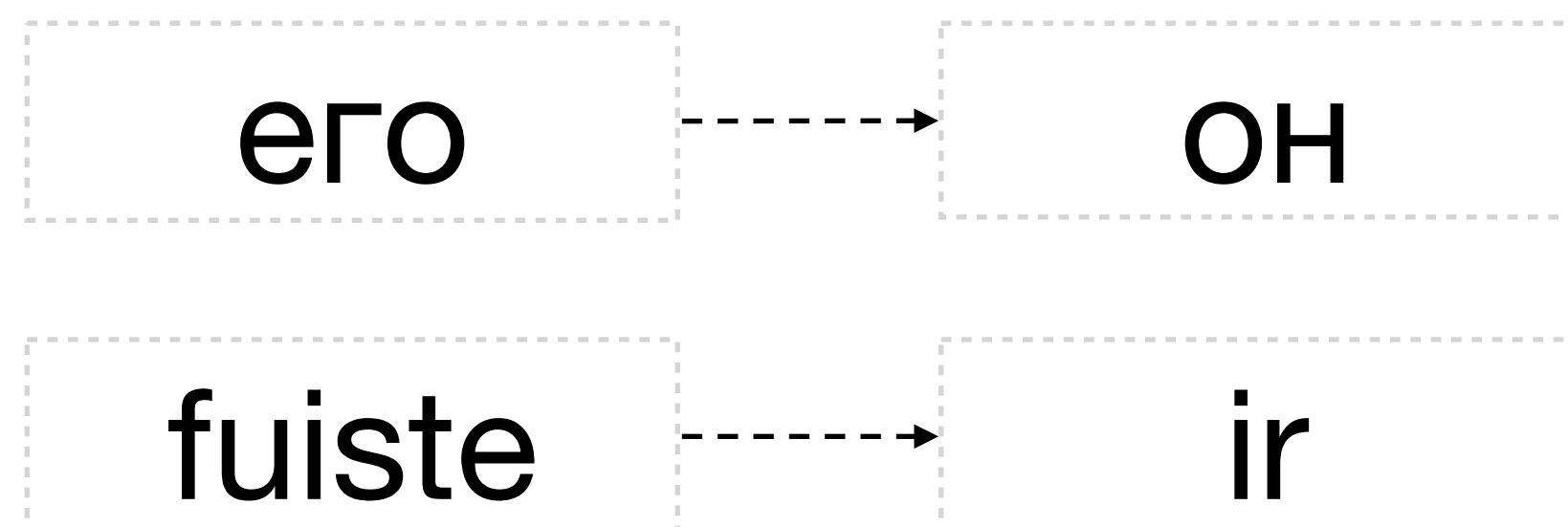
- Символы слова
- Морфологические признаки
- ...

Классификаторы:

- CNN
- Perceptron
- ...

Нормализация (лемматизация) слов

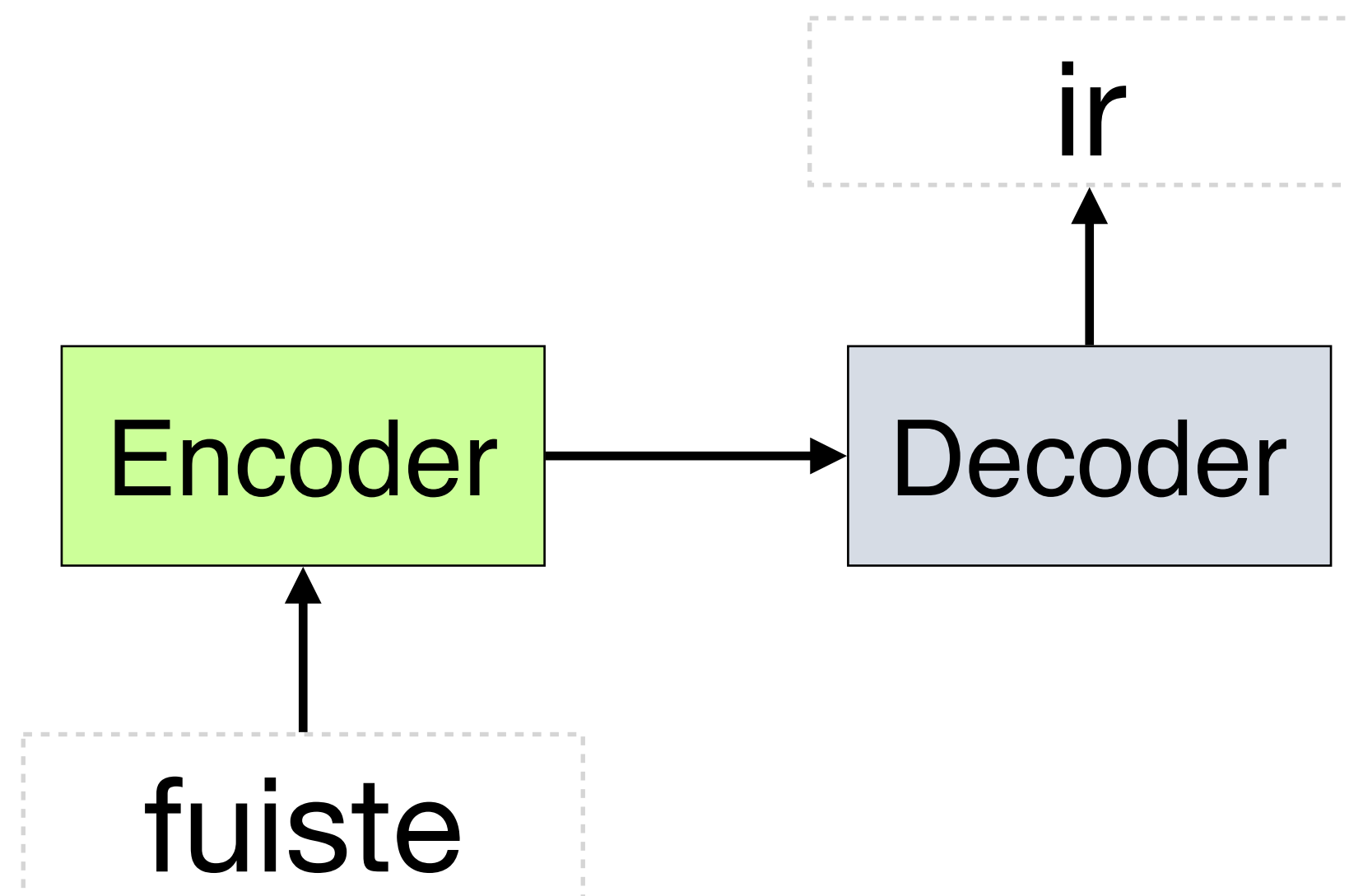
- Важная проблема таких подходов — слова исключения



- Решение — словари исключений

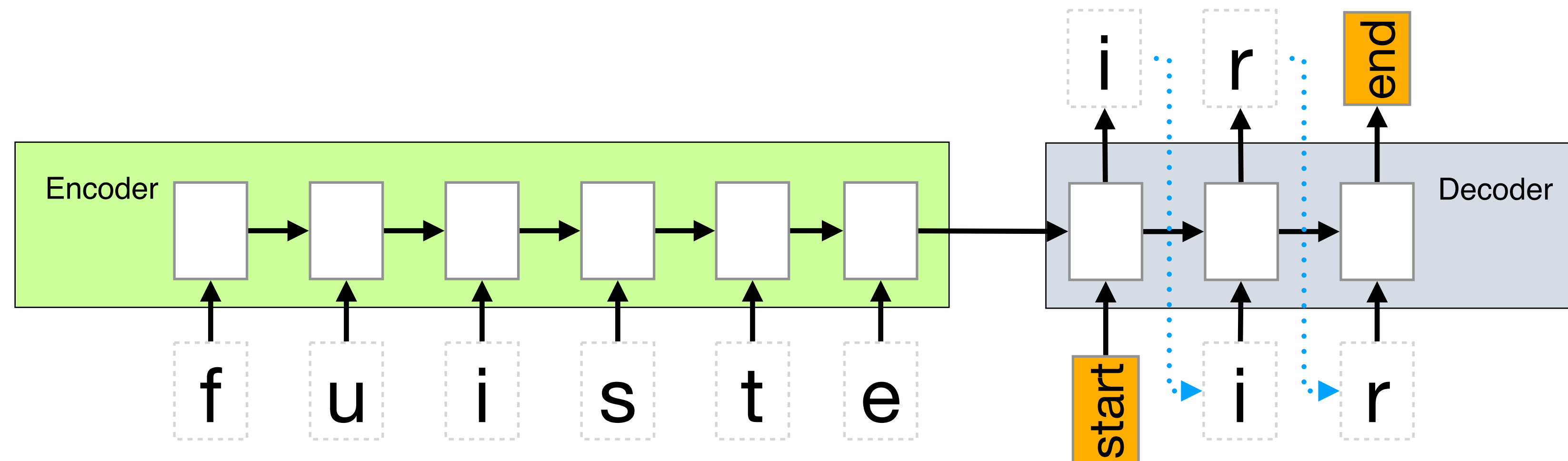
Нормализация (лемматизация) слов

seq2seq



Нормализация (лемматизация) слов

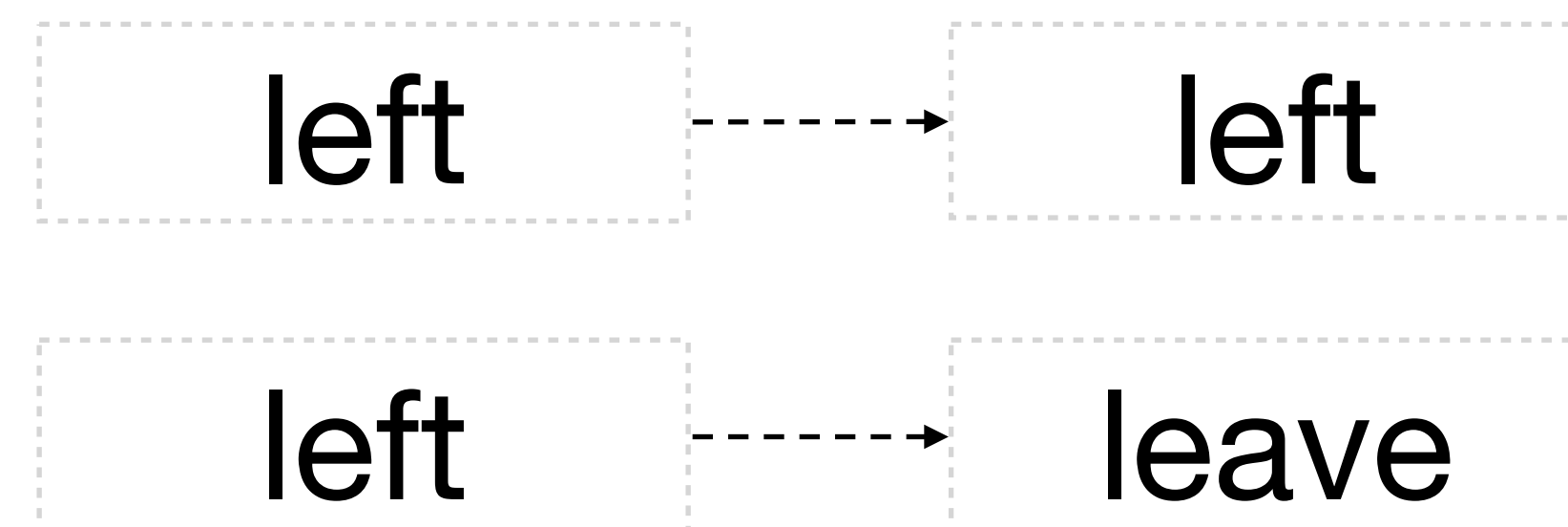
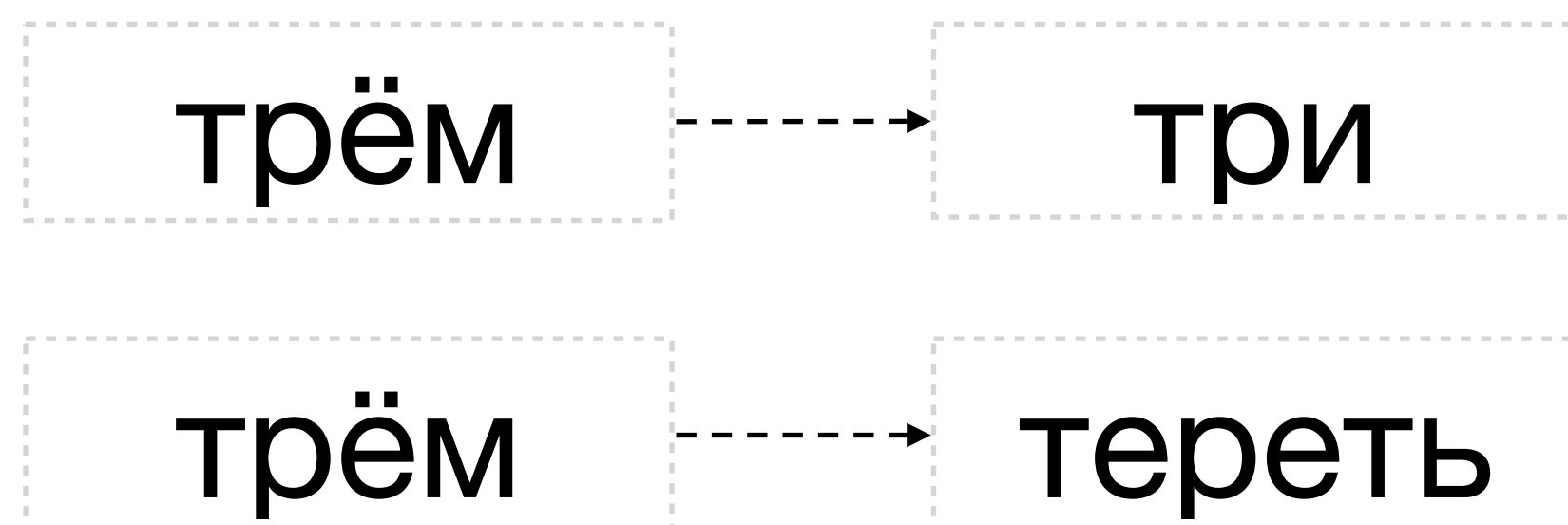
seq2seq



Нормализация (лемматизация) слов

Многозначность

Грамматическая омонимия



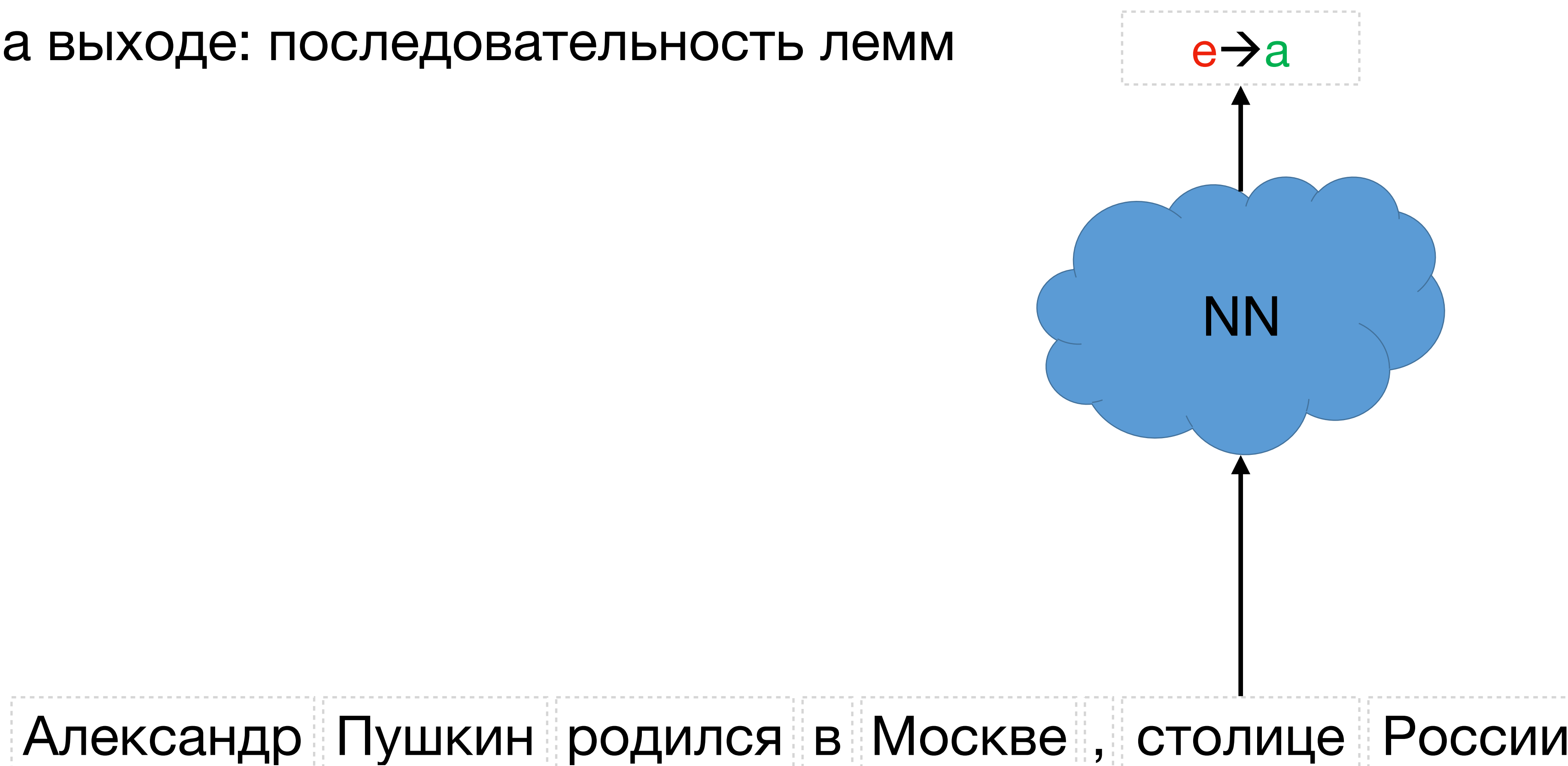
Решение

- Использовать информацию о морфологии слова

Нормализация (лемматизация) слов

Задача

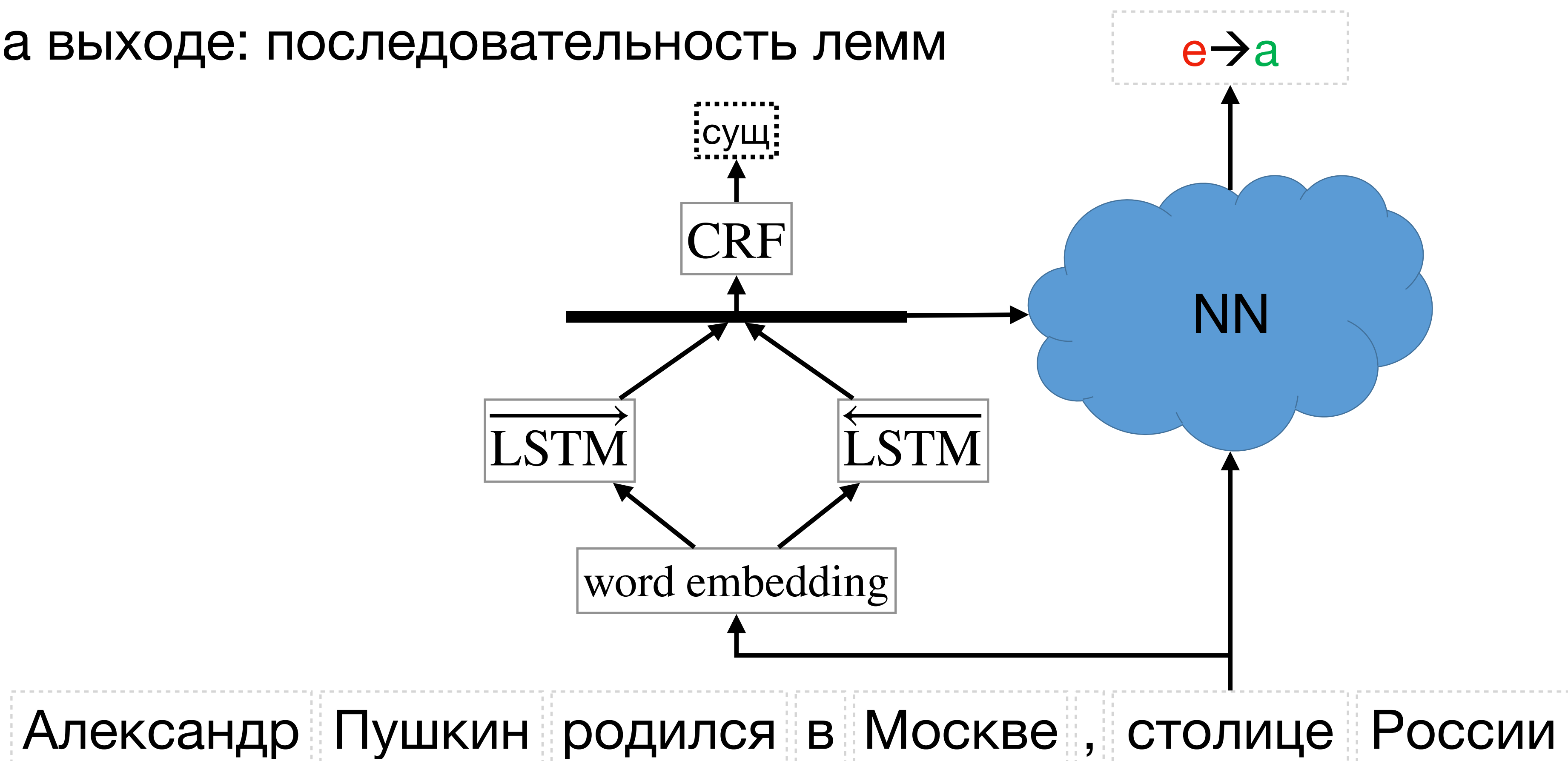
- На входе: последовательность слов (токенов)
- На выходе: последовательность лемм



Нормализация (лемматизация) слов

Задача

- На входе: последовательность слов (токенов)
- На выходе: последовательность лемм



Нормализация (лемматизация) слов

Оценка качества

- Классическая точность (Accuracy)

$$\text{Accuracy} = \frac{\text{correct}}{\text{total}}$$

- Точность среди неизвестных слов (out-of-vocabulary)

$$\text{Accuracy}_{\text{OOV}} = \frac{\text{correct}_{\text{OOV}}}{\text{total}_{\text{OOV}}}$$

Следующая лекция

Синтаксический анализ

- Дерево составляющих
 - Алгоритм СКУ
- Дерево зависимостей
 - transition-based parsing
 - графовые методы