Part-of-speech tagging с использованием нейронных сетей

Даниил Анастасьев

Научный руководитель: Евгений Инденбом

Москва, 2018

Содержание

Описание задачи

- Part-of-speech tagging — важный источник признаков для большинства NLP pipeline'ов.

Описание задачи

- Part-of-speech tagging — важный источник признаков для большинства NLP pipeline'ов.
- Задача — найти грамматические значения (*теги*) всех слов в предложении:

У двери стоял	**стол** NOUN Animacy=Inan Case=Nom Gender=Masc Number=Sing	секретарши , ⋯
The interviews took	**place** NN	two years ago .

Омонимичность тегов

Грамматическое значение слова почти невозможно определить, не принимая во внимание его контекст:

she	hated	lies
	VBD	
PRP	VBN	NNS
	JJ	VBZ

Обзор поставленной задачи

- Чтобы решить любую задачу машинного обучения, нам нужны:

Обзор поставленной задачи

- Чтобы решить любую задачу машинного обучения, нам нужны:
  1. Модель;

Обзор поставленной задачи

- Чтобы решить любую задачу машинного обучения, нам нужны:
  1. Модель;
  2. Данные для обучения;

Обзор поставленной задачи

- Чтобы решить любую задачу машинного обучения, нам нужны:
  1. Модель;
  2. Данные для обучения;
  3. Признаки, извлекаемые из данных;

Обзор поставленной задачи

- Чтобы решить любую задачу машинного обучения, нам нужны:
  1. Модель;
  2. Данные для обучения;
  3. Признаки, извлекаемые из данных;
  4. Функция потерь.

Обзор поставленной задачи

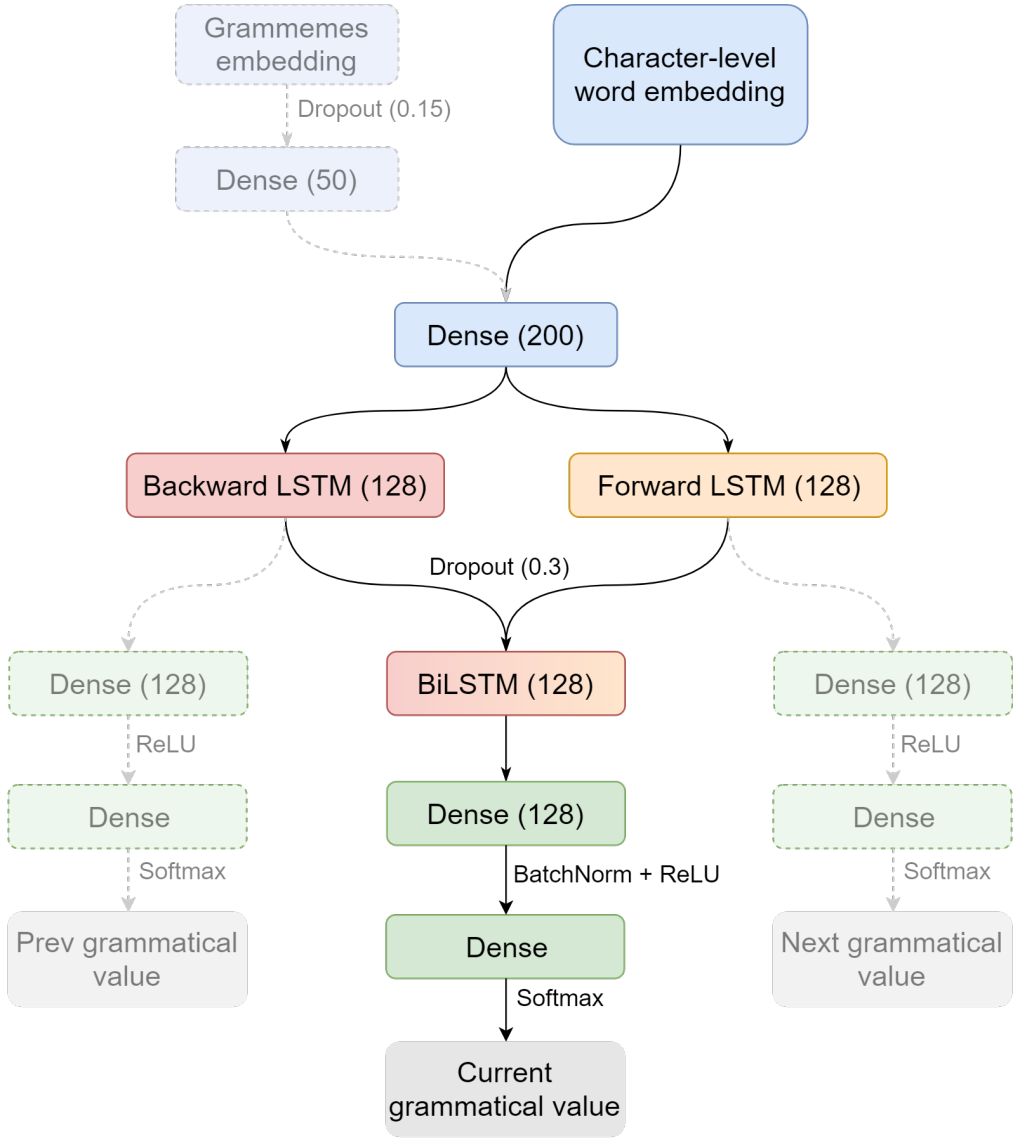
- Чтобы решить любую задачу машинного обучения, нам нужны:
  1. Модель;
  2. Данные для обучения;
  3. Признаки, извлекаемые из данных;
  4. Функция потерь.
- Зафиксируем модель — Bidirectional LSTM;

Обзор поставленной задачи

- Чтобы решить любую задачу машинного обучения, нам нужны:
  1. Модель;
  2. Данные для обучения;
  3. Признаки, извлекаемые из данных;
  4. Функция потерь.
- Зафиксируем модель — Bidirectional LSTM;
- Улучшим её результат, сконцентрировавшись на остальных компонентах.

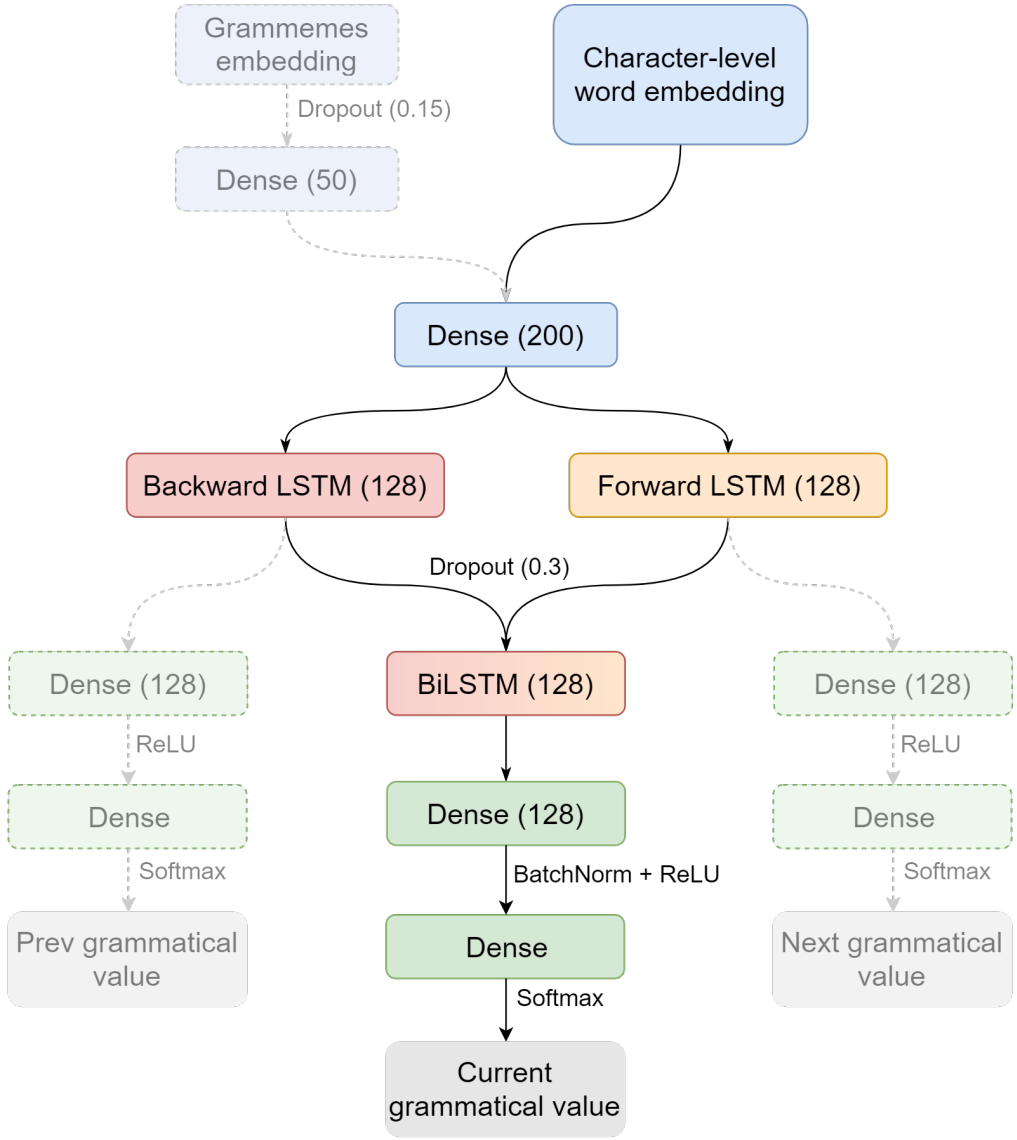
Baseline

- Основой модели для нас послужит двуслойный BiLSTM;



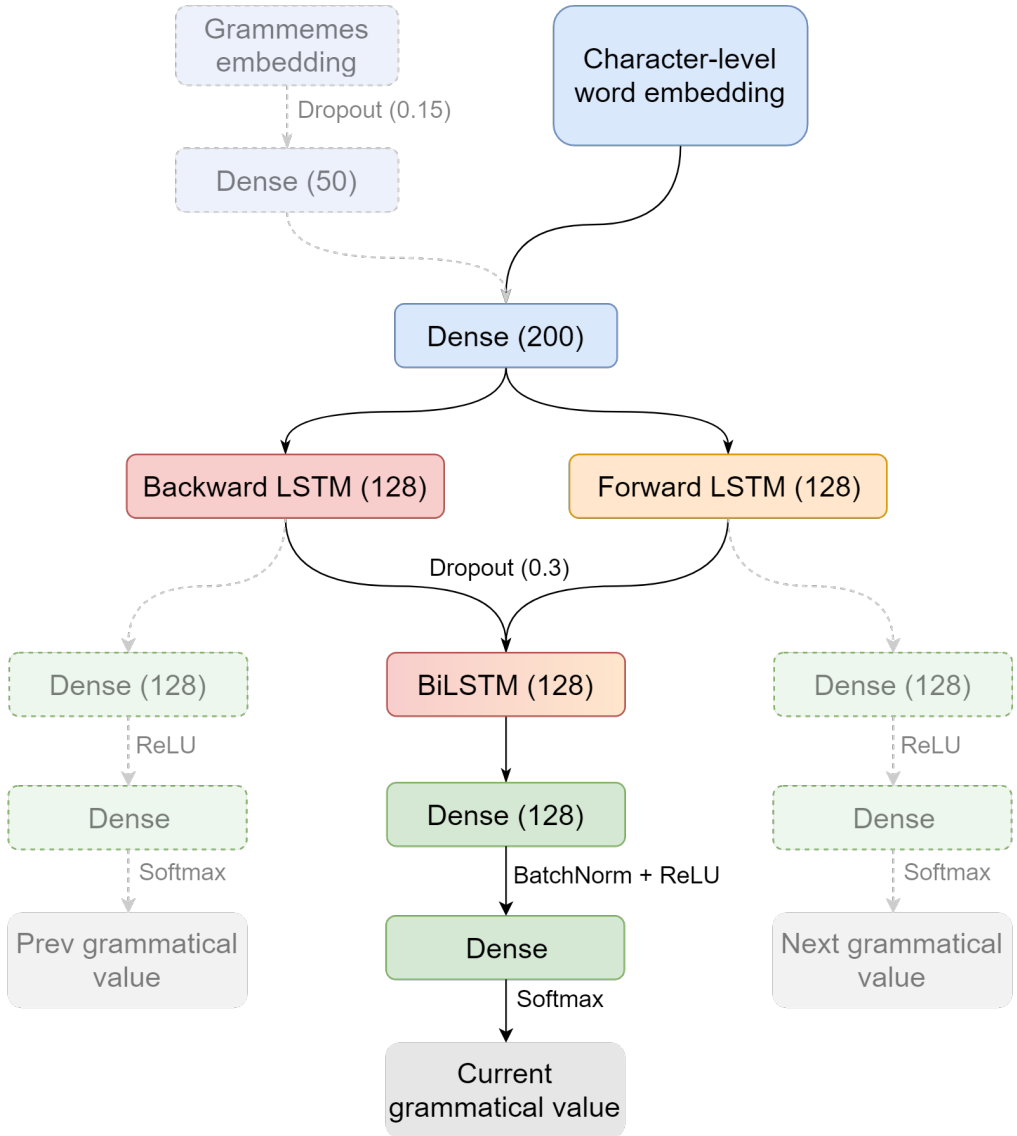
Baseline

- Основой модели для нас послужит двуслойный BiLSTM;
- Эмбеддинги слов будет строить BiLSTM символьного уровня — часть модели с самым высоким качеством на PTB;



Baseline

- Основой модели для нас послужит двуслойный BiLSTM;
- Эмбеддинги слов будет строить BiLSTM символьного уровня — часть модели с самым высоким качеством на PTB;
- Дополним пошагово данную модель новыми фишками.



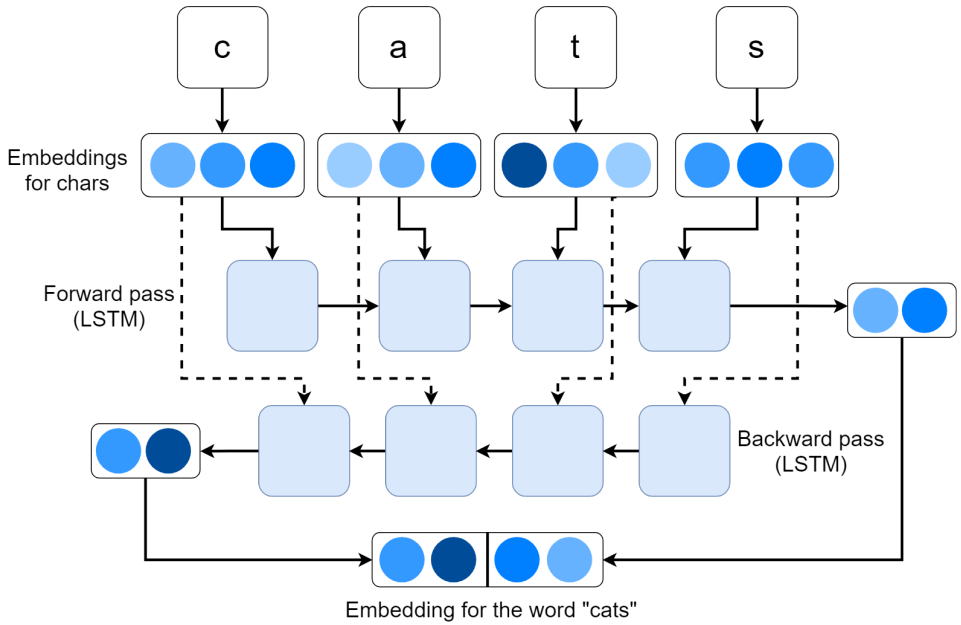
Датасеты

Использовались следующие датасеты для проверки моделей:

Датасет	**Train**	**Dev**	**Test**	**#labels**
PTB	912 344	131 768	129 654	45
UD SynTagRus	871 082	118 630	117 470	723
MorphoRuEval-2017	977 567	108 581	19 560	302
Tiger	711 041	88 152	89 054	54
UD Ukrainian IU	75 098	10 371	14 939	1 196

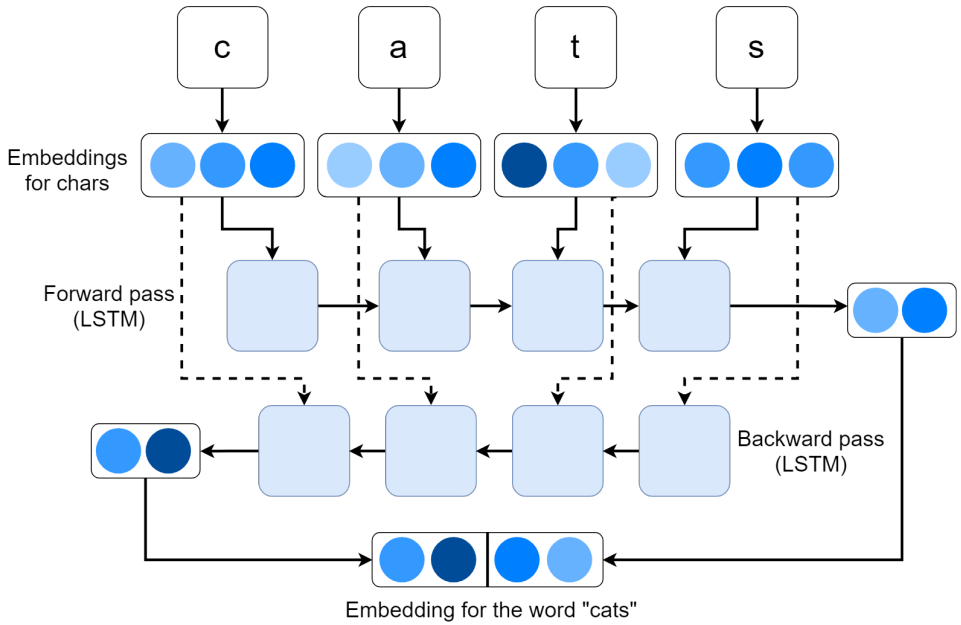
Char BiLSTM

- BiLSTM символьного уровня — один из стандартных способов построить словный эмбеддинг;



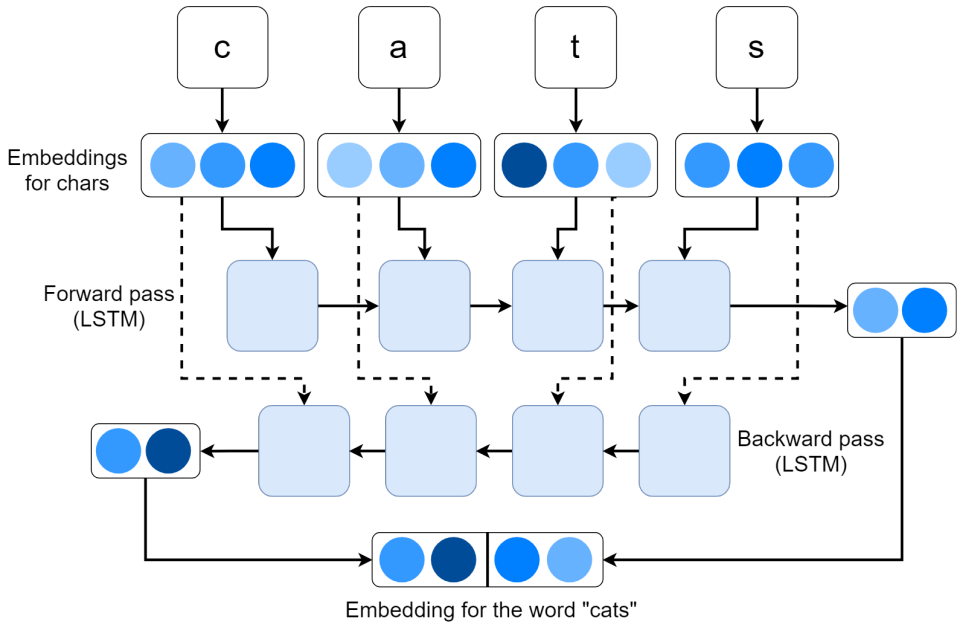
Char BiLSTM

- BiLSTM символьного уровня — один из стандартных способов построить словный эмбеддинг;
- Обрабатывает символы один за другим;



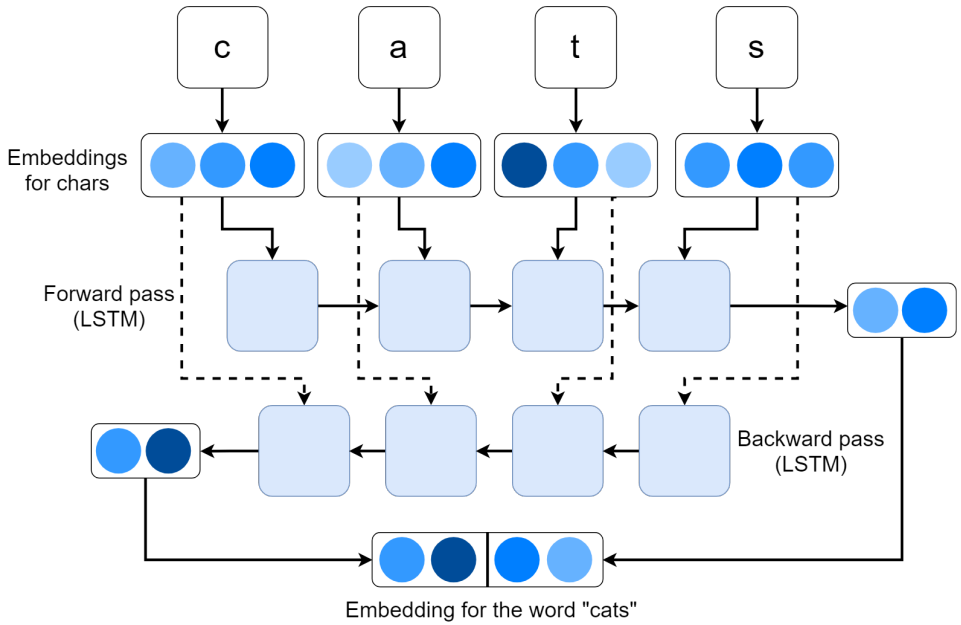
Char BiLSTM

- BiLSTM символьного уровня — один из стандартных способов построить словный эмбеддинг;
- Обрабатывает символы один за другим;
- Может обрабатывать слова произвольной длины;



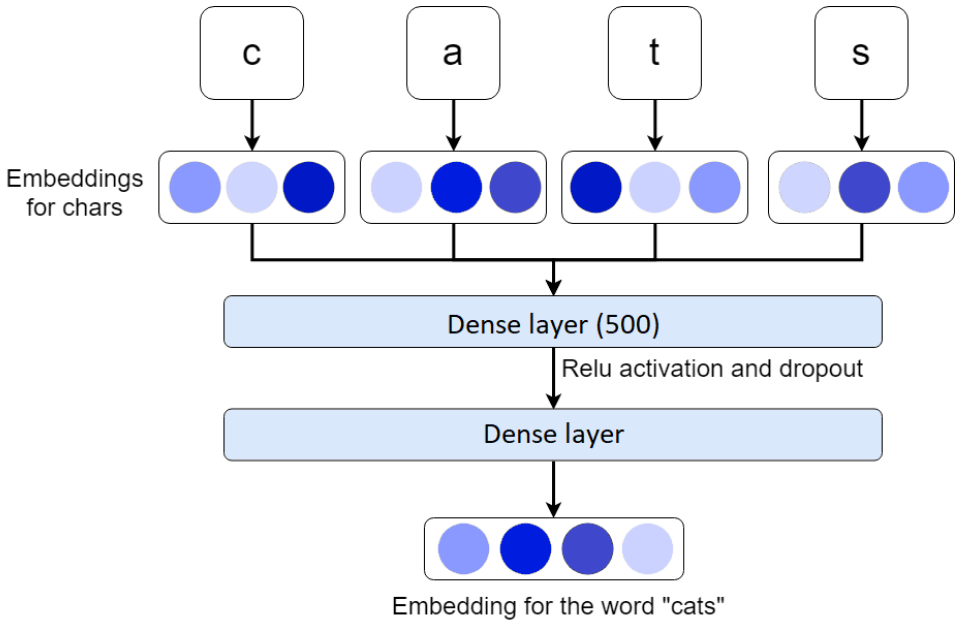
Char BiLSTM

- BiLSTM символьного уровня — один из стандартных способов построить словный эмбеддинг;
- Обрабатывает символы один за другим;
- Может обрабатывать слова произвольной длины;
- Не параллелизуется эффективно.



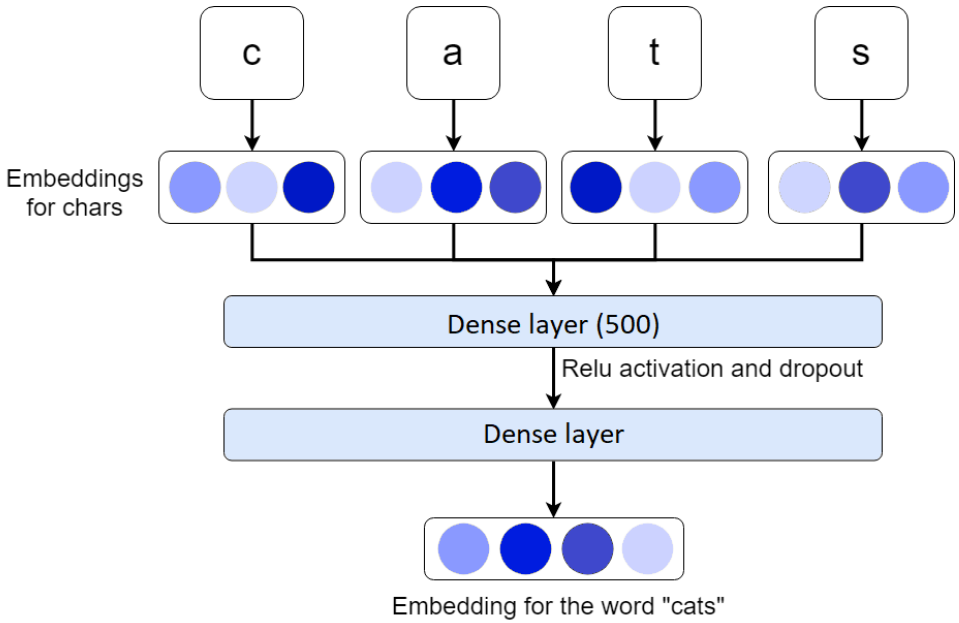
Char FF

- Обычная feed-forward сеть — предлагаемая альтернатива Char BiLSTM;



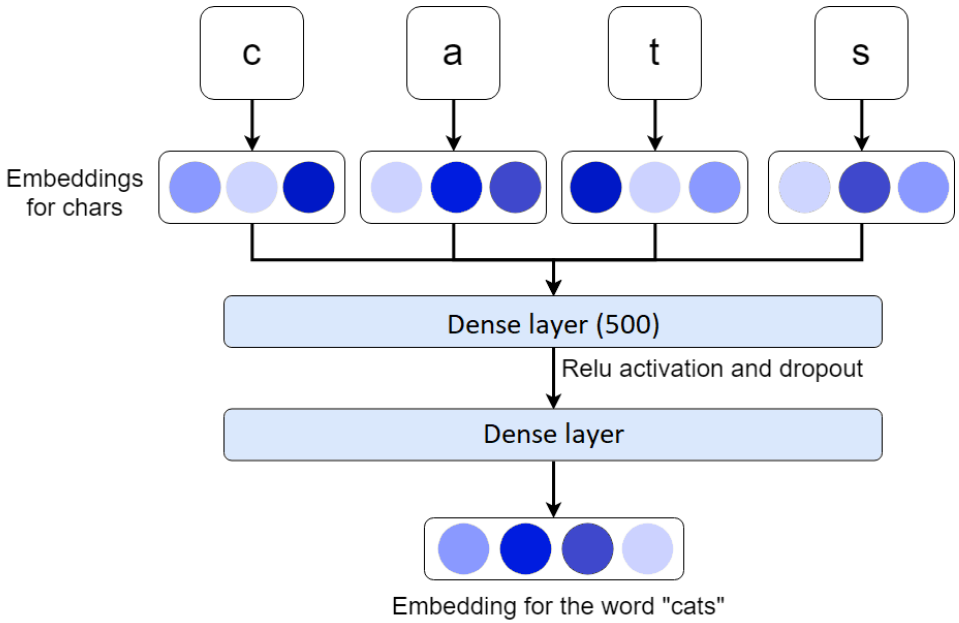
Char FF

- Обычная feed-forward сеть — предлагаемая альтернатива Char BiLSTM;
- Обрабатывает конкатенацию символьных эмбеддингов;



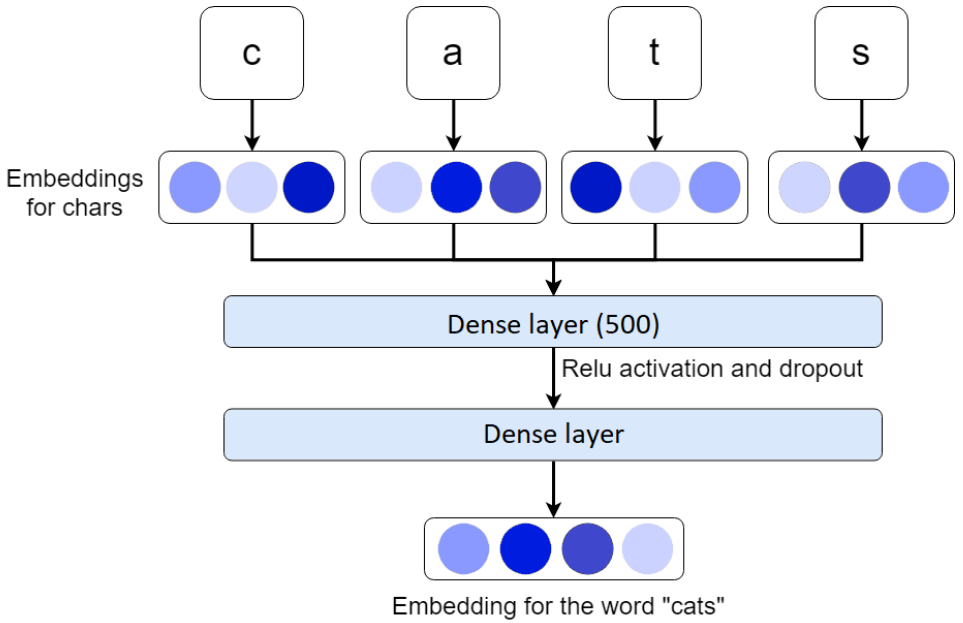
Char FF

- Обычная feed-forward сеть — предлагаемая альтернатива Char BiLSTM;
- Обрабатывает конкатенацию символьных эмбеддингов;
- Может работать лишь с фиксированной длиной слов:



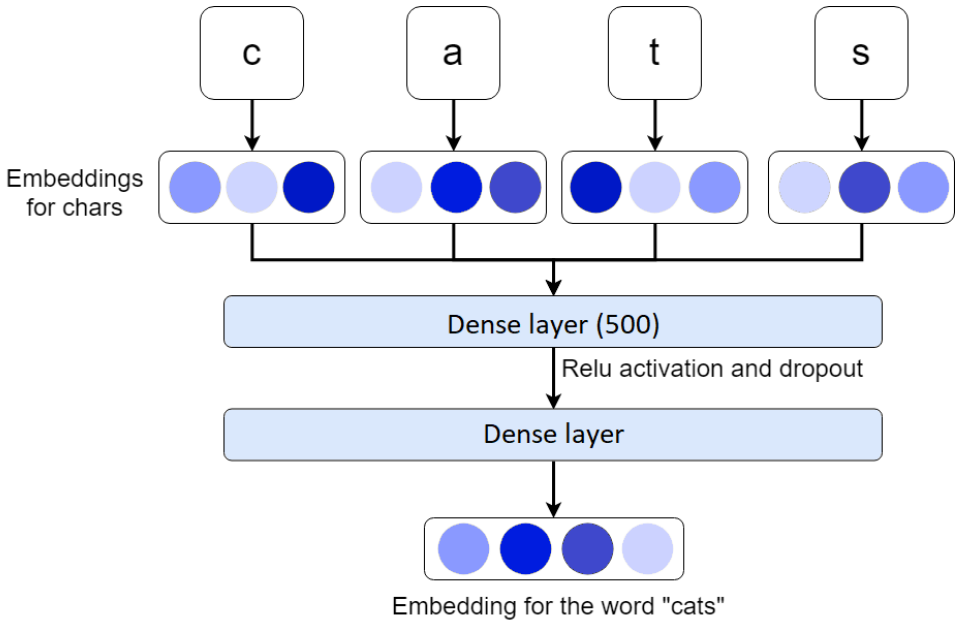
Char FF

- Обычная feed-forward сеть — предлагаемая альтернатива Char BiLSTM;
- Обрабатывает конкатенацию символьных эмбеддингов;
- Может работать лишь с фиксированной длиной слов:
  - 11–13 символов обычно достаточно;



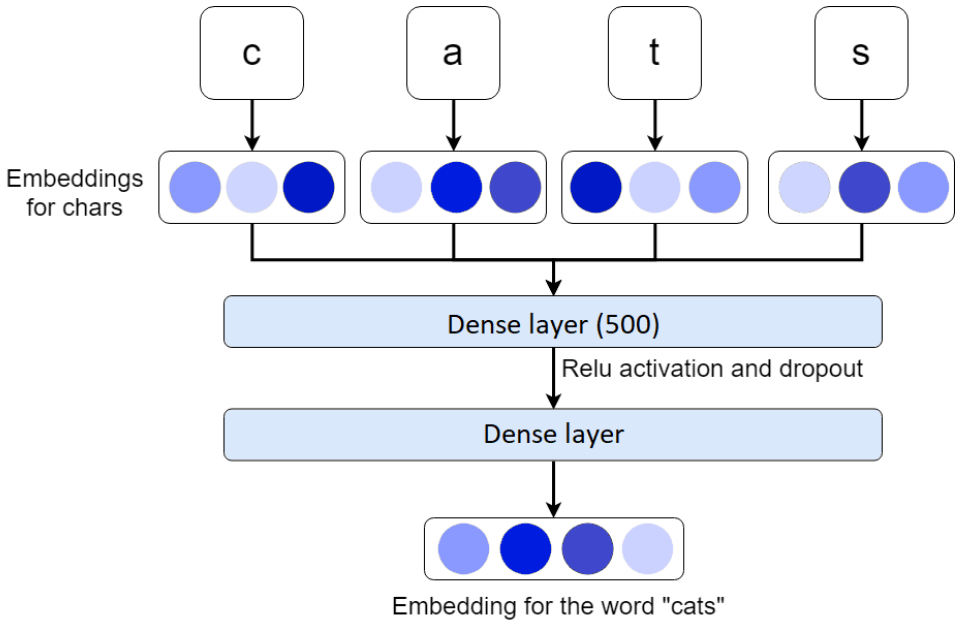
Char FF

- Обычная feed-forward сеть — предлагаемая альтернатива Char BiLSTM;
- Обрабатывает конкатенацию символьных эмбеддингов;
- Может работать лишь с фиксированной длиной слов:
  - 11–13 символов обычно достаточно;
  - Более короткие слова дополняются слева, более длинные обрезаются.



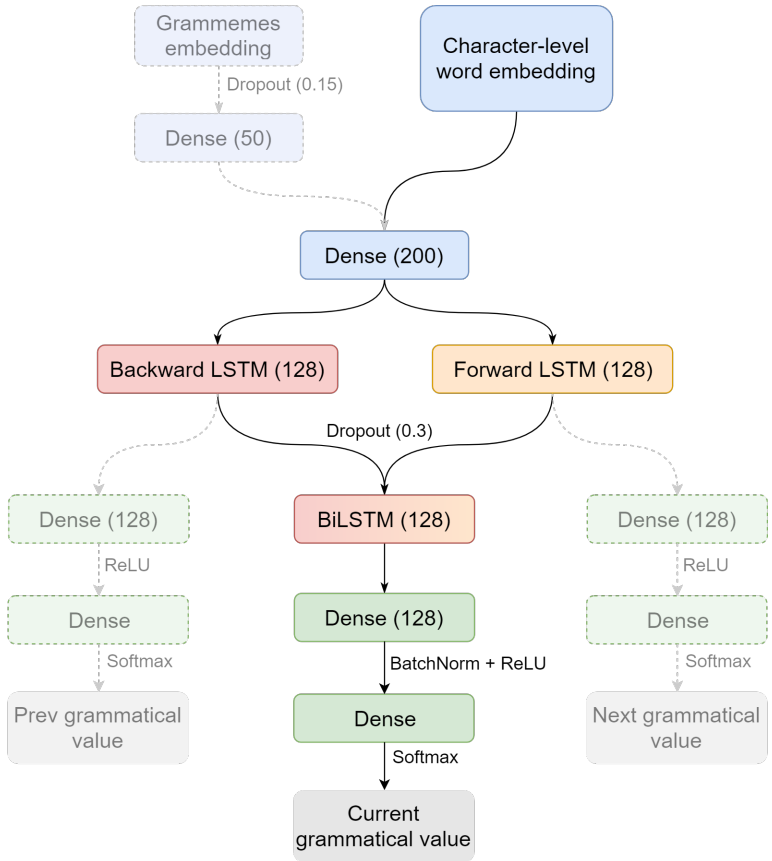
Char FF

- Обычная feed-forward сеть — предлагаемая альтернатива Char BiLSTM;
- Обрабатывает конкатенацию символьных эмбеддингов;
- Может работать лишь с фиксированной длиной слов:
  - 11–13 символов обычно достаточно;
  - Более короткие слова дополняются слева, более длинные обрезаются.
- Считаются гораздо быстрее Char BiLSTM.



Сравнение вариантов эмбеддингов

- По качеству модели близки;
- По скорости очевидно выигрывает Char FF.



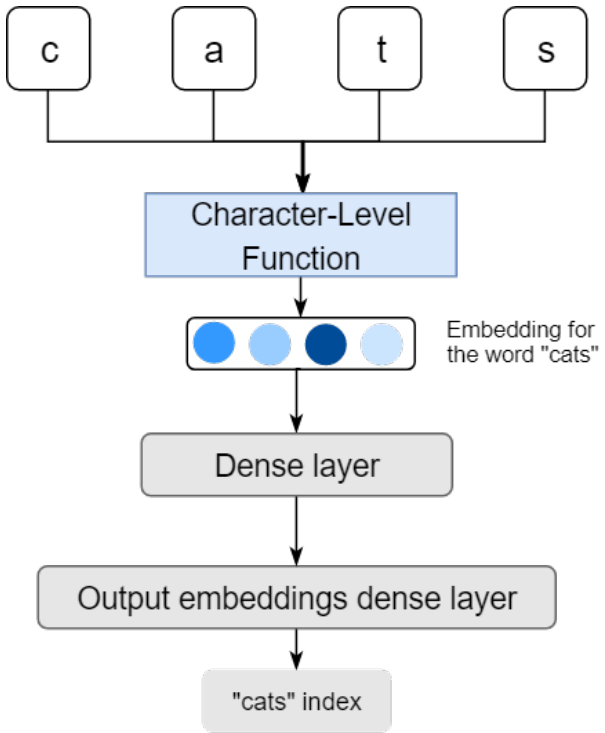
Dataset	Char BiLSTM	Char FF
PTB	97.02% / 96.98%	**97.32%** / **97.26%**
SynTagRus	**95.23%** / **95.39%**	94.98% / 95.16%
MorphoRuEval	96.48% / **94.69%**	**96.68%** / 94.63%
Tiger	98.27% / **99.86%**	**98.31%** / 99.73%
Ukrainian	80.10% / 78.70%	**81.51%** / **79.48%**

Предобучение эмбеддингов

- В предобученных словных эмбеддингах закодирована важная информация, собранная на больших корпусах;

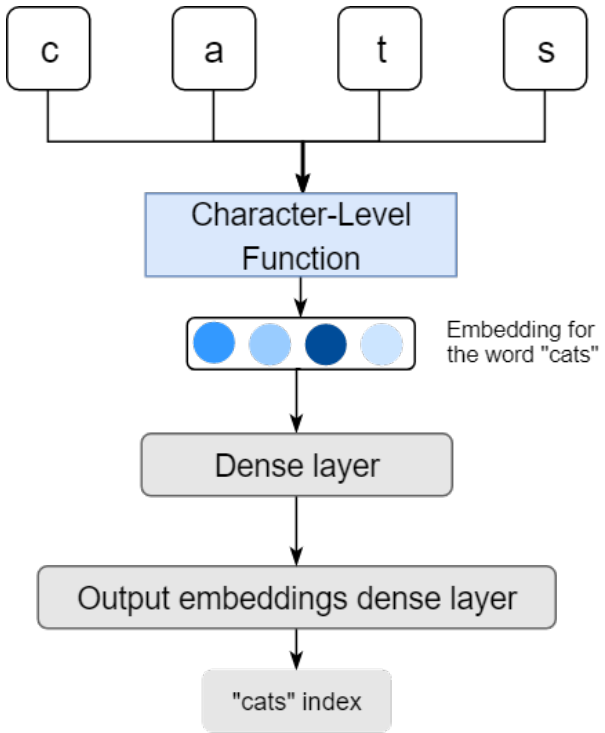
Предобучение эмбеддингов

- В предобученных словных эмбеддингах закодирована важная информация, собранная на больших корпусах;
- Перенесём эту информацию с помощью автоэнкодера:



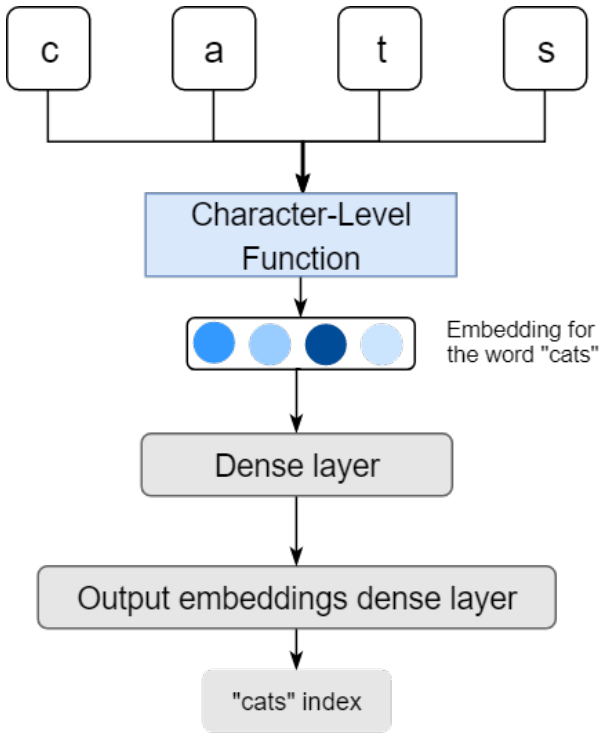
Предобучение эмбеддингов

- В предобученных словных эмбеддингах закодирована важная информация, собранная на больших корпусах;
- Перенесём эту информацию с помощью автоэнкодера:
  - Энкодер — одна из функций символьного уровня;



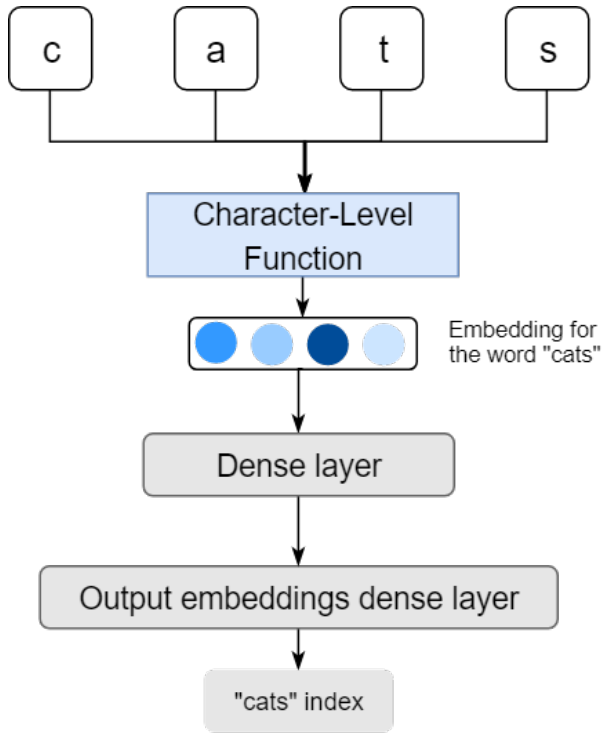
Предобучение эмбеддингов

- В предобученных словных эмбеддингах закодирована важная информация, собранная на больших корпусах;
- Перенесём эту информацию с помощью автоэнкодера:
  - Энкодер — одна из функций символьного уровня;
  - Декодер — полносвязный слой, инициализированный предобученными эмбеддингами.



Предобучение эмбеддингов

- В предобученных словных эмбеддингах закодирована важная информация, собранная на больших корпусах;
- Перенесём эту информацию с помощью автоэнкодера:
  - Энкодер — одна из функций символьного уровня;
  - Декодер — полносвязный слой, инициализированный предобученными эмбеддингами.
- Кросс-энтропийные потери стимулируют символьный эмбеддинг слова приближаться по косинусной мере к его предобученному варианту и удаляться от всех остальных эмбеддингов.



Результаты при применении предобучения

- Предобученные эмбеддинги обучались дальше вместе со всей моделью под задачу;

Результаты при применении предобучения

- Предобученные эмбеддинги обучались дальше вместе со всей моделью под задачу;
- На первых эпохах модель с предобучением достигала заметно более высокого качества.

Результаты при применении предобучения

- Предобученные эмбеддинги обучались дальше вместе со всей моделью под задачу;
- На первых эпохах модель с предобучением достигала заметно более высокого качества.

Dataset	**Char FF**	**Char FF (Pretrained)**
PTB	97.32% / 97.26%	**97.40%** / **97.31%**
SynTagRus	94.98% / 95.16%	**95.22%** / **95.36%**
MorphoRuEval	96.68% / 94.63%	**96.88%** / **94.63%**
Tiger	98.31% / **99.73%**	**98.39%** / 99.69%
Ukrainian	81.51% / 79.48%	**82.65%** / **80.67%**

Граммемные эмбеддинги

ABBYY

- Сложно предсказать тег слова по набору его символов — близкие по написанию слова бывают очень далеки;

Граммемные эмбеддинги

ABBYY

- Сложно предсказать тег слова по набору его символов — близкие по написанию слова бывают очень далеки;
- Будем оценивать априорные вероятности каждой из возможных граммем по словарю:

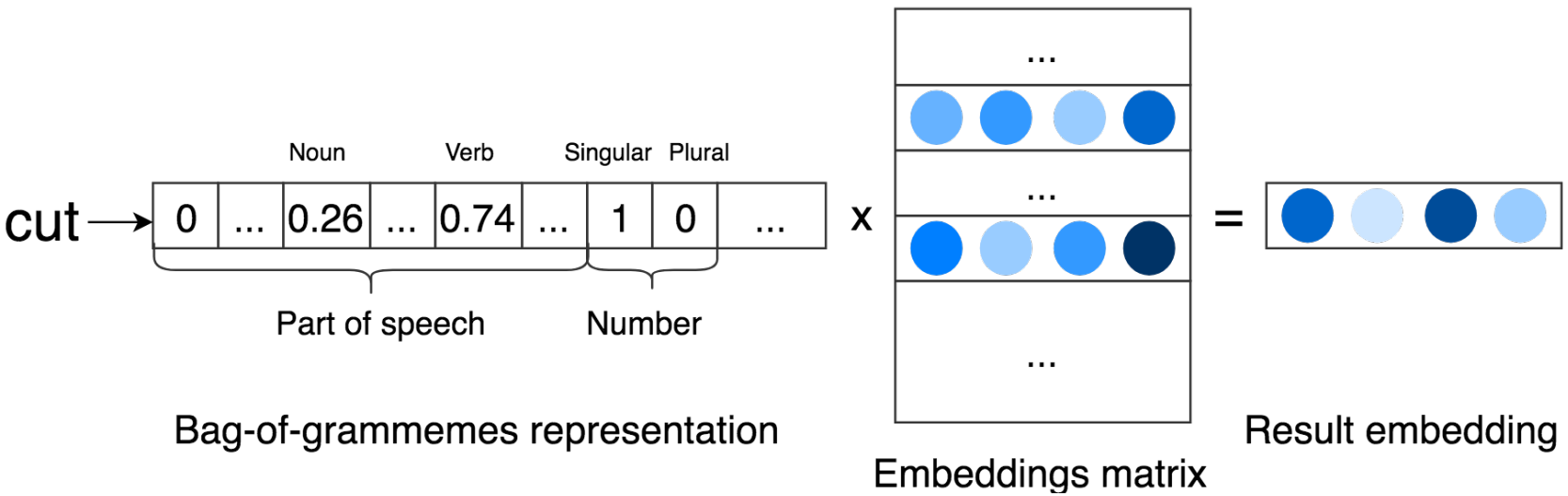
Граммемные эмбеддинги

ABBYY

- Сложно предсказать тег слова по набору его символов — близкие по написанию слова бывают очень далеки;
- Будем оценивать априорные вероятности каждой из возможных граммем по словарю:
  - Например, форма существительного «cut» имеет частотность $2.84 \cdot 10^{-5}$, глагольная форма — $8.75 \cdot 10^{-5}$.
    Тогда $\mathbf{P}(\text{noun}) \approx 0.26$.

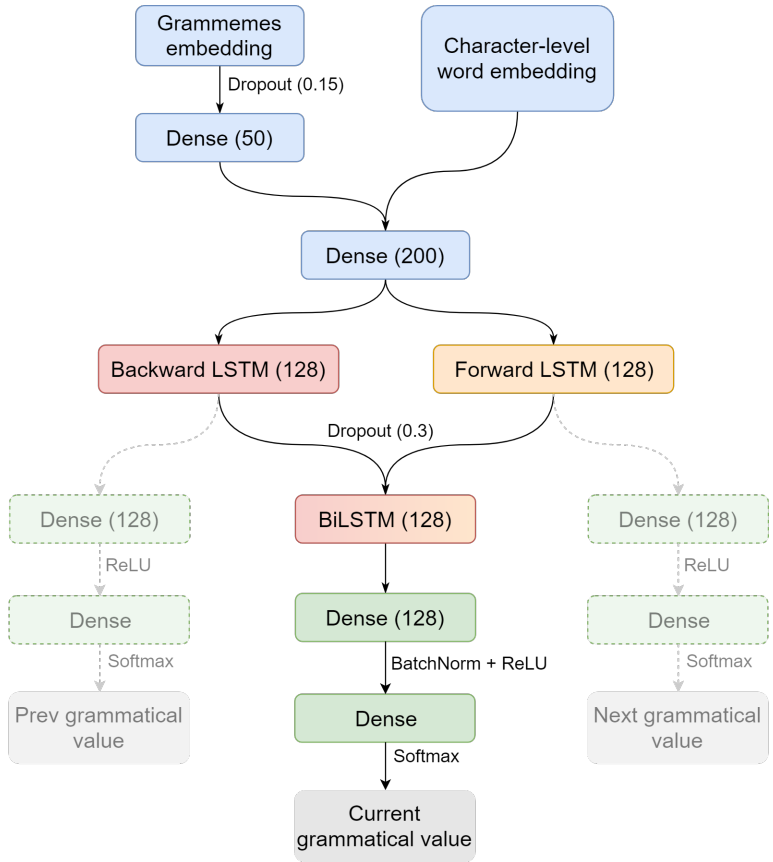
Граммемные эмбеддинги

- Сложно предсказать тег слова по набору его символов — близкие по написанию слова бывают очень далеки;
- Будем оценивать априорные вероятности каждой из возможных граммем по словарю:
  - Например, форма существительного «cut» имеет частотность $2.84 \cdot 10^{-5}$, глагольная форма — $8.75 \cdot 10^{-5}$.
    Тогда $\mathbf{P}(\text{noun}) \approx 0.26$.
- Добавим полносвязный слой для сокращения размерности эмбеддингов:



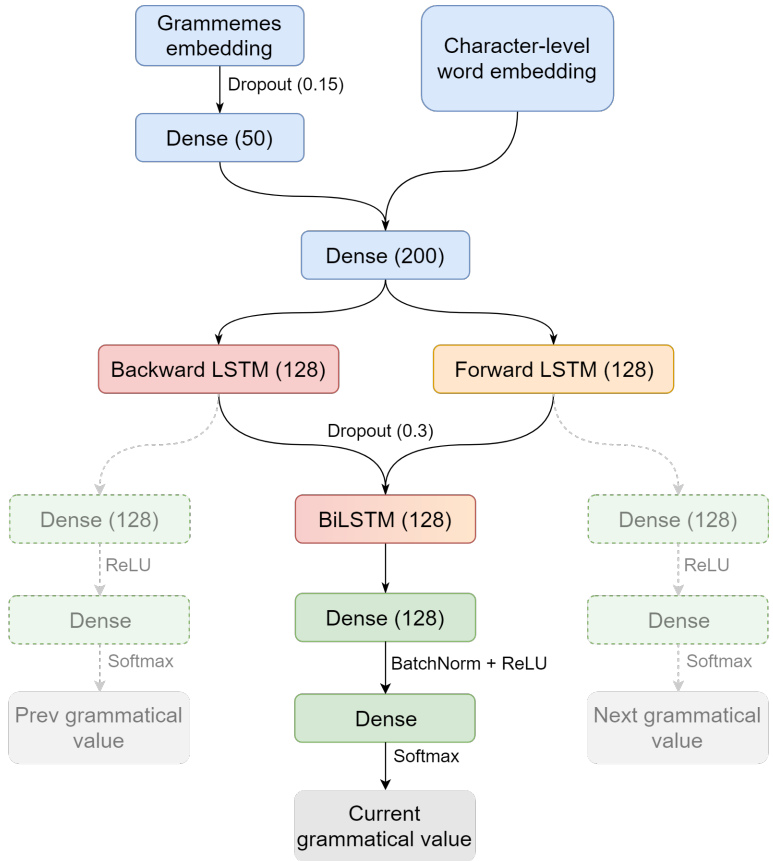
Результаты при применении граммемных эмбеддингов

- На русском и украинском наиболее заметный прирост — до 35–43% ERR;
- На английском и немецком прирост незначительный.



Результаты при применении граммемных эмбеддингов

- На русском и украинском наиболее заметный прирост — до 35–43% ERR;
- На английском и немецком прирост незначительный.



Dataset	Char FF (Pretrained)	+ Grammemes
PTB	97.40% / **97.31%**	**97.43%** / 97.30%
SynTagRus	95.22% / 95.36%	**96.77%** / **97.00%**
MorphoRuEval	96.88% / 94.63%	**98.07%** / **95.36%**
Tiger	98.39% / 99.69%	**98.70%** / **99.85%**
Ukrainian	82.65% / 80.67%	**89.61%** / **88.06%**

Результаты при добавлении языкового моделирования

- Добавим в модель потери от языкового моделирования;

Результаты при добавлении языкового моделирования

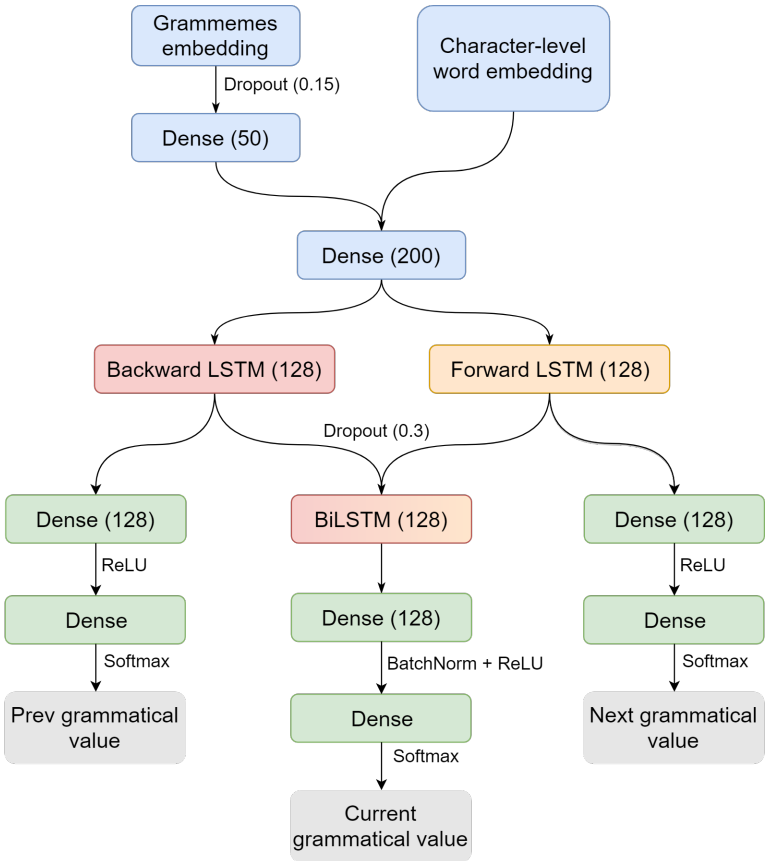
- Добавим в модель потери от языкового моделирования;
- *POS LM* пытается вместе с предсказанием тега слова выдавать теги предыдущего и следующего слов:

$$\text{Forward LSTM(she, hated)} \sim \begin{matrix}\text{tag(hated)} \\ \text{tag(lies)}\end{matrix}$$

Результаты при добавлении языкового моделирования

- Добавим в модель потери от языкового моделирования;
- *POS LM* пытается вместе с предсказанием тега слова выдавать теги предыдущего и следующего слов:

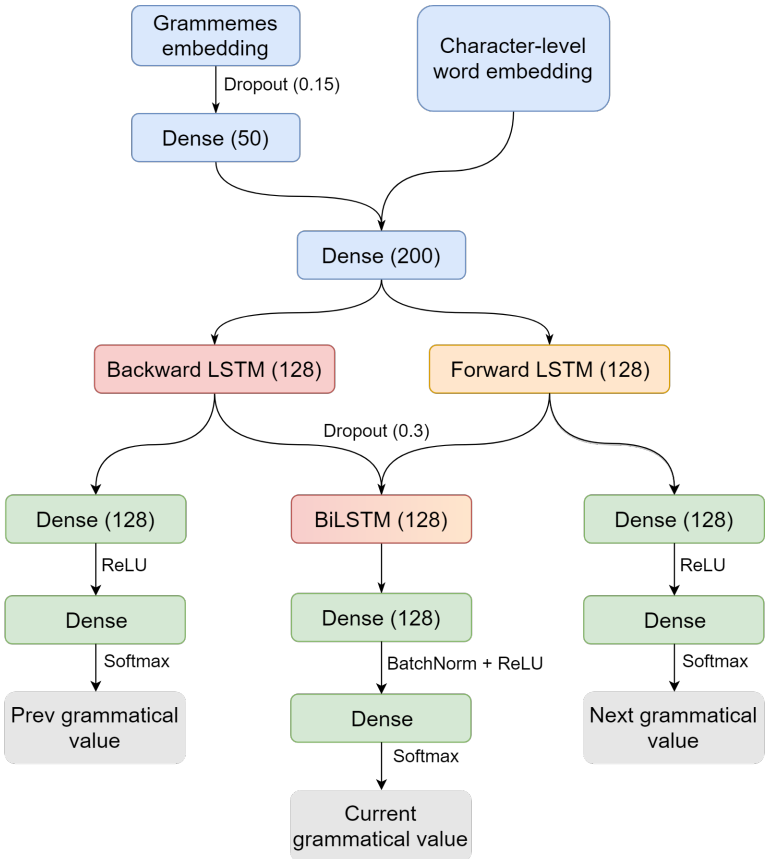
$$\text{Forward LSTM(she, hated)} \sim \begin{matrix}\text{tag(hated)} \\ \text{tag(lies)}\end{matrix}$$



Результаты при добавлении языкового моделирования

- Добавим в модель потери от языкового моделирования;
- *POS LM* пытается вместе с предсказанием тега слова выдавать теги предыдущего и следующего слов:

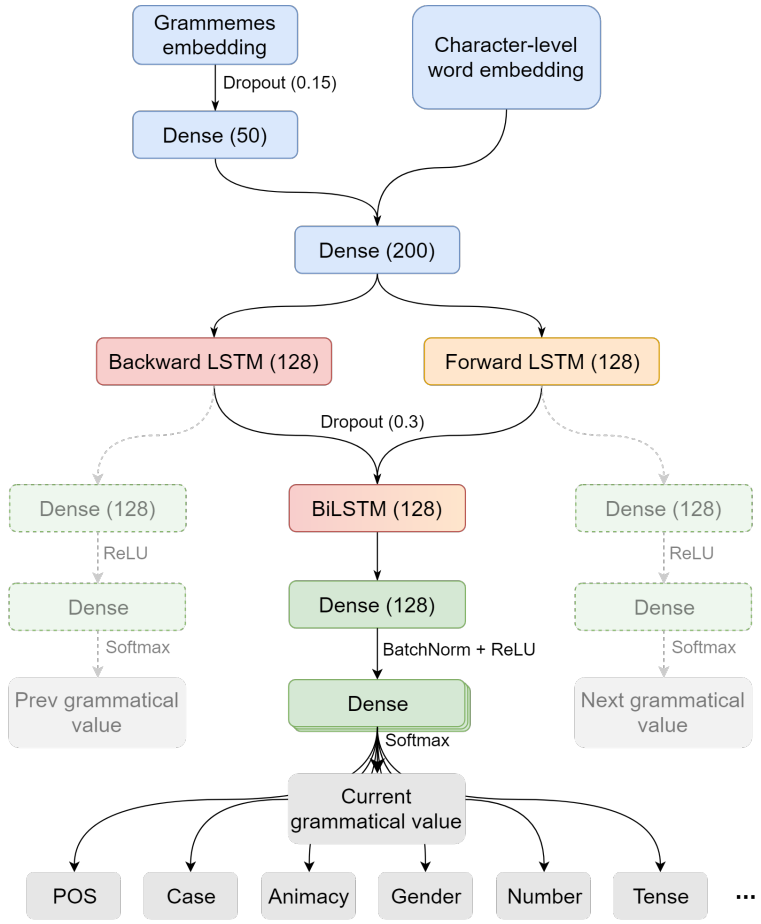
Forward LSTM(she, hated) $\sim$ tag(hated) tag(lies)



Dataset	All features	+ POS LM
PTB	97.43% / 98.30%	**97.57%** / **97.49%**
SynTagRus	96.77% / 97.00%	**96.97%** / **97.24%**
MorphoRuEval	98.07% / 94.85%	**98.12%** / **96.72%**
Tiger	98.70% / **99.85%**	**98.71%** / 99.57%
Ukrainian	**89.61%** / 88.06%	89.48% / **88.07%**

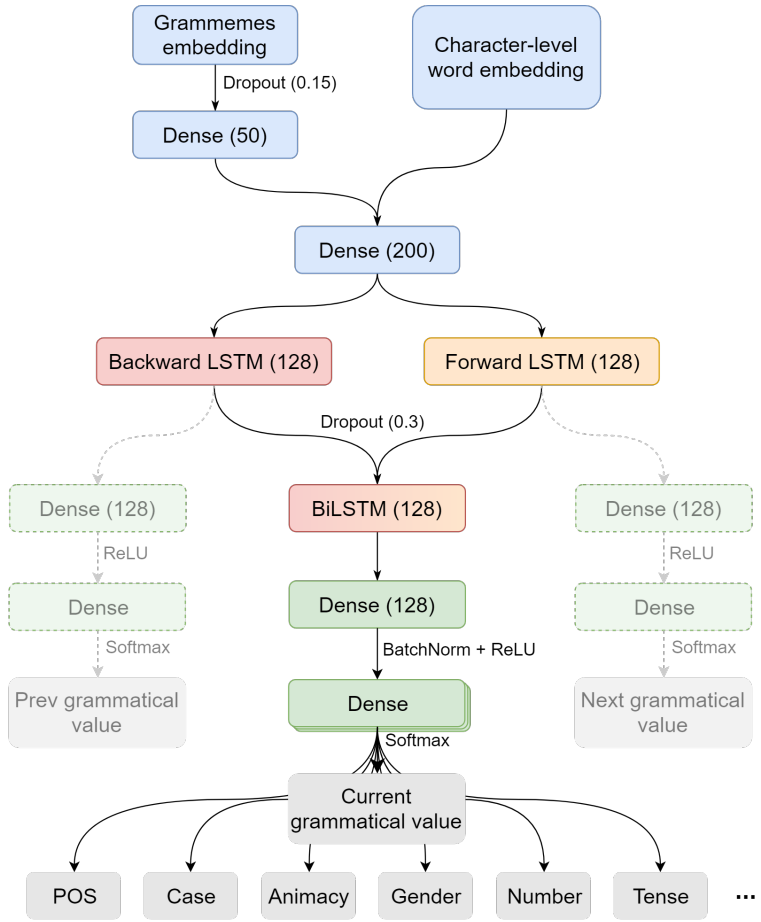
Предсказание отдельных граммем

- Добавим в модель предсказание отдельных граммем для каждой из возможных грамматических категорий:



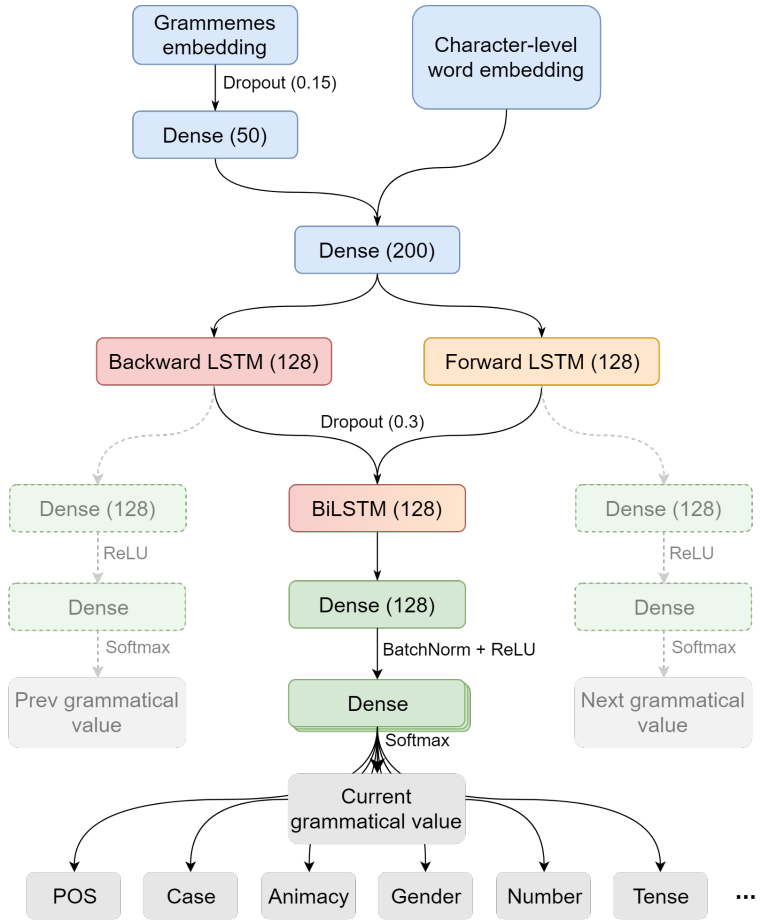
Предсказание отдельных граммем

- Добавим в модель предсказание отдельных граммем для каждой из возможных грамматических категорий:
  - Например, для категории Number будем предсказывать распределение на граммемах Single, Plural и Undefined.



Предсказание отдельных граммем

- Добавим в модель предсказание отдельных граммем для каждой из возможных грамматических категорий:
  - Например, для категории Number будем предсказывать распределение на граммемах Single, Plural и Undefined.



Dataset	All features	+ POS LM	+ Gram categories
SynTagRus	96.77% / 97.00%	**96.97%** / **97.24%**	96.89% / 97.20%
MorphoRuEval	98.07% / 94.85%	**98.12%** / **96.72%**	98.01% / 96.65%
Ukrainian	89.61% / 88.06%	89.48% / 88.07%	**90.17%** / **89.01%**

Перенос модели между датасетами

- Перенесем модель на SynTagRus с
  1. MorphoRuEval датасета с похожим UD тегсетом;
  2. Размеченным Compreno датасетом с не слишком похожим тегсетом.
- При переносе заменим выходной слой и будем первые несколько эпох тренировать только его, а только затем — всю модель целиком.

Перенос модели между датасетами

- Перенесем модель на SynTagRus с
  1. MorphoRuEval датасета с похожим UD тегсетом;
  2. Размеченным Compreno датасетом с не слишком похожим тегсетом.
- При переносе заменим выходной слой и будем первые несколько эпох тренировать только его, а только затем — всю модель целиком.

Model	**Accuracy**
Best previous	96.97% / 97.24%
MorphoRuEval pretrained	**98.21%** / **98.33%**
Compreno pretrained	98.18% / 98.29%

Перенос модели между языками

- Перенесем модель на украинский язык с SynTagRus аналогичным образом;

Перенос модели между языками

- Перенесем модель на украинский язык с SynTagRus аналогичным образом;
- Для этого объединим все граммемы и символы, встречающиеся в этих языках.

Dataset	**Char FF + Grammemes + PosLM**	**+ Pretrained on Syntagrus**
Ukrainian	89.48% / 88.07%	**90.93%** / **89.54%**

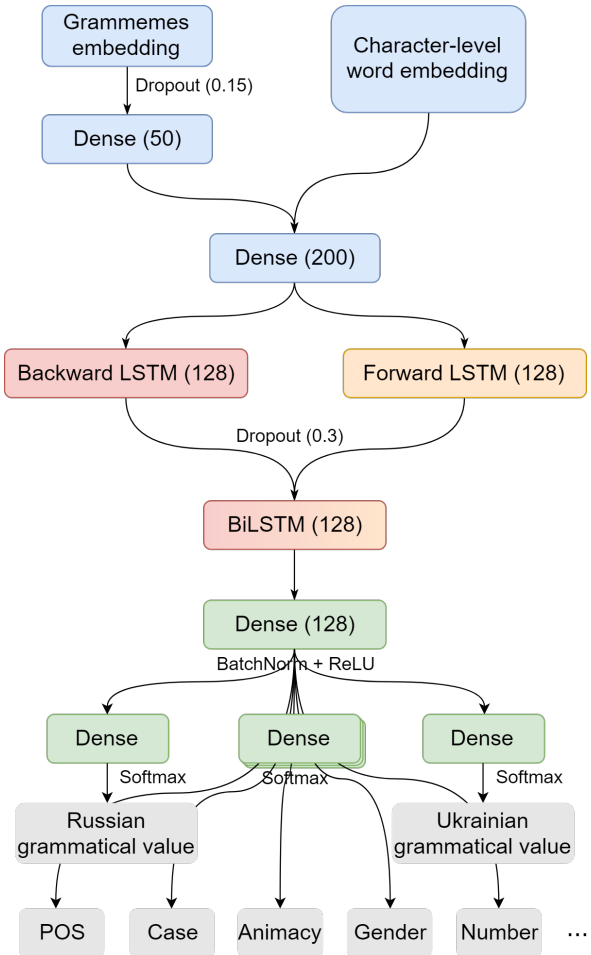
Совместная тренировка модели под несколько языков

ABBYY

- Модель при переносе очень быстро забывает то, на чём она училась до этого;

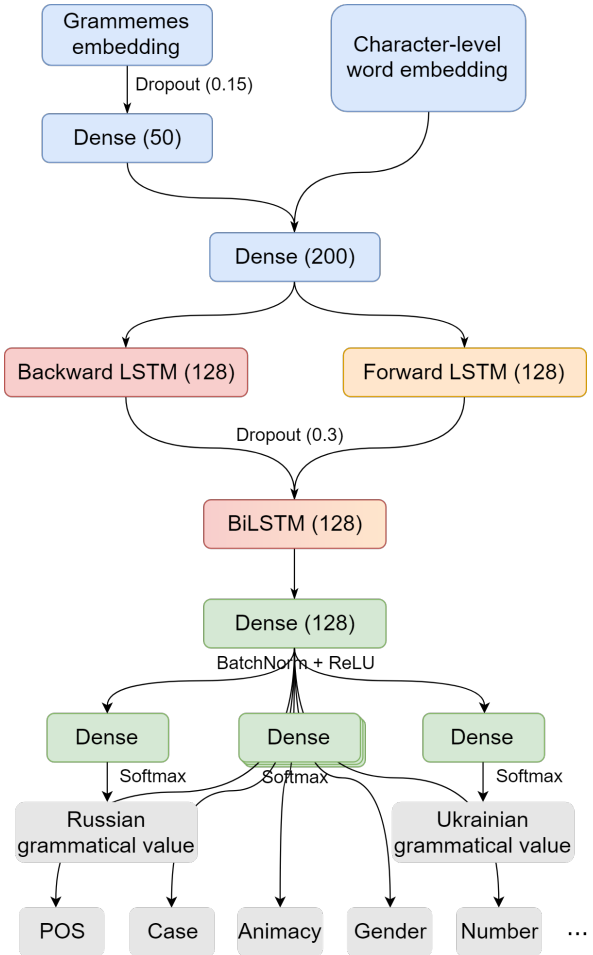
Совместная тренировка модели под несколько языков

- Модель при переносе очень быстро забывает то, на чём она училась до этого;
- Будем тренировать модель под два языка сразу;



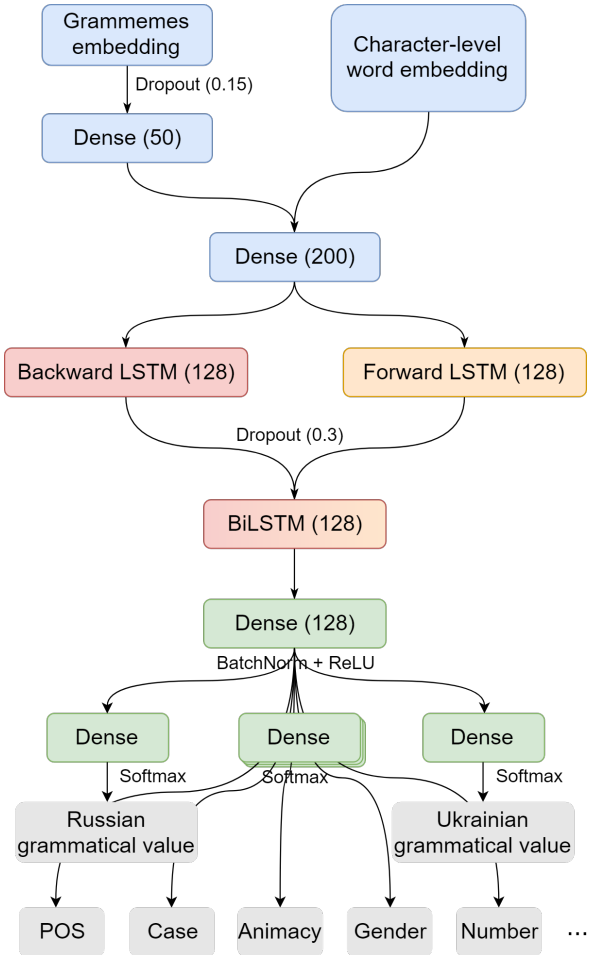
Совместная тренировка модели под несколько языков

- Модель при переносе очень быстро забывает то, на чём она училась до этого;
- Будем тренировать модель под два языка сразу;
- Добавим к модели предсказание отдельных граммем;



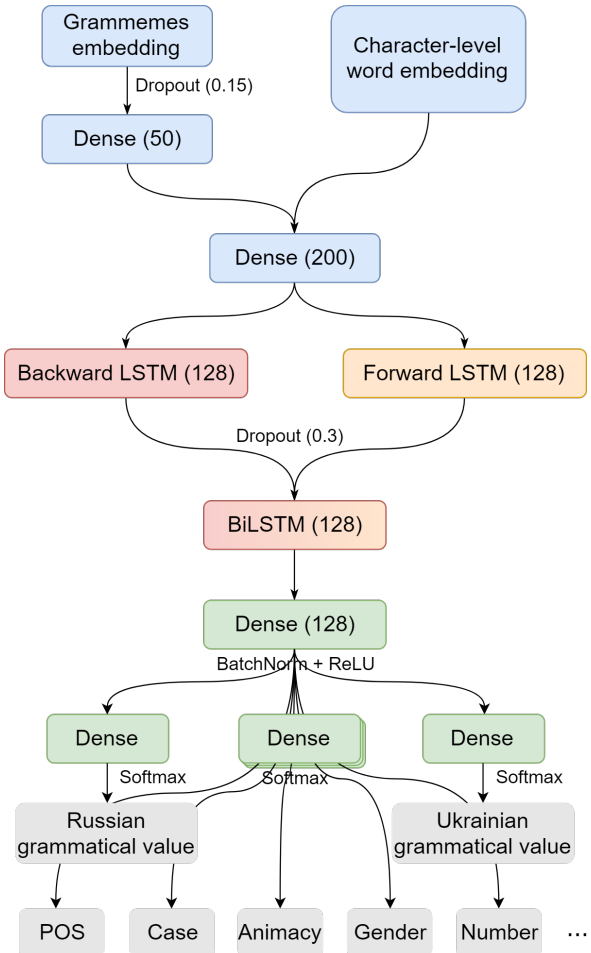
Совместная тренировка модели под несколько языков

- Модель при переносе очень быстро забывает то, на чём она училась до этого;
- Будем тренировать модель под два языка сразу;
- Добавим к модели предсказание отдельных граммем;
- Слои предсказания граммем будут не language specific.



Совместная тренировка модели под несколько языков

- Модель при переносе очень быстро забывает то, на чём она училась до этого;
- Будем тренировать модель под два языка сразу;
- Добавим к модели предсказание отдельных граммем;
- Слои предсказания граммем будут не language specific.



Dataset	Transfer baseline	Multi-lang	Multi-lang + Gram categories
Ukrainian	90.93% / 89.54%	91.15% / 89.33%	**91.72%** / **89.87%**
Syntagrus	**96.77%** / 97.00%	96.44% / 96.69%	96.66% / **97.01%**

Сравнение с baseline

- С использованием всех улучшений оказывается возможным значительно превзойти baseline;

Dataset	**Char BiLSTM**	**Best Model**	**ERR**
PTB	97.02% / 96.98%	**97.60%** / **97.51%**	19.4% / 17.5%
SynTagRus	95.23% / 95.39%	**98.21%** / **98.33%**	62.5% / 63.8%
MorphoRuEval	96.48% / 94.69%	**98.12%** / **96.72%**	46.5% / 38.2%
Tiger	98.27% / 99.86%	**98.74%** / **99.91%**	27.2% / 35.7%
Ukrainian	80.10% / 78.70%	**91.72%** / **89.87%**	58.4% / 52.4%

Сравнение с baseline

- С использованием всех улучшений оказывается возможным значительно превзойти baseline;
- При этом размер модели почти не вырос;

Dataset	**Char BiLSTM**	**Best Model**	**ERR**
PTB	97.02% / 96.98%	**97.60%** / **97.51%**	19.4% / 17.5%
SynTagRus	95.23% / 95.39%	**98.21%** / **98.33%**	62.5% / 63.8%
MorphoRuEval	96.48% / 94.69%	**98.12%** / **96.72%**	46.5% / 38.2%
Tiger	98.27% / 99.86%	**98.74%** / **99.91%**	27.2% / 35.7%
Ukrainian	80.10% / 78.70%	**91.72%** / **89.87%**	58.4% / 52.4%

Сравнение с baseline

- С использованием всех улучшений оказывается возможным значительно превзойти baseline;
- При этом размер модели почти не вырос;
- Итоговая модель значительно меньше большинства state-of-the-art моделей, но показывает сопоставимое качество.

Dataset	**Char BiLSTM**	**Best Model**	**ERR**
PTB	97.02% / 96.98%	**97.60%** / **97.51%**	19.4% / 17.5%
SynTagRus	95.23% / 95.39%	**98.21%** / **98.33%**	62.5% / 63.8%
MorphoRuEval	96.48% / 94.69%	**98.12%** / **96.72%**	46.5% / 38.2%
Tiger	98.27% / 99.86%	**98.74%** / **99.91%**	27.2% / 35.7%
Ukrainian	80.10% / 78.70%	**91.72%** / **89.87%**	58.4% / 52.4%

Выводы

- Мы начали с сильного baseline — BiLSTM модель с Char BiLSTM эмбеддингами;

Выводы

- Мы начали с сильного baseline — BiLSTM модель с Char BiLSTM эмбеддингами;
- Был предложен более быстрый аналог Char BiLSTM, показывающий сопоставимое качество;

Выводы

- Мы начали с сильного baseline — BiLSTM модель с Char BiLSTM эмбеддингами;
- Был предложен более быстрый аналог Char BiLSTM, показывающий сопоставимое качество;
- Был разработан метод для предобучения эмбеддингов символьного уровня;

Выводы

- Мы начали с сильного baseline — BiLSTM модель с Char BiLSTM эмбеддингами;
- Был предложен более быстрый аналог Char BiLSTM, показывающий сопоставимое качество;
- Был разработан метод для предобучения эмбеддингов символьного уровня;
- Были описаны дополнительные функции потерь, улучшающие качество целевой функции — POS LM и предсказание отдельных граммем;

Выводы

- Мы начали с сильного baseline — BiLSTM модель с Char BiLSTM эмбеддингами;
- Был предложен более быстрый аналог Char BiLSTM, показывающий сопоставимое качество;
- Был разработан метод для предобучения эмбеддингов символьного уровня;
- Были описаны дополнительные функции потерь, улучшающие качество целевой функции — POS LM и предсказание отдельных граммем;
- Был продемонстрирован положительный эффект от переноса модели с датасета на датасет и с языка на язык;

Выводы

- Мы начали с сильного baseline — BiLSTM модель с Char BiLSTM эмбеддингами;
- Был предложен более быстрый аналог Char BiLSTM, показывающий сопоставимое качество;
- Был разработан метод для предобучения эмбеддингов символьного уровня;
- Были описаны дополнительные функции потерь, улучшающие качество целевой функции — POS LM и предсказание отдельных граммем;
- Был продемонстрирован положительный эффект от переноса модели с датасета на датасет и с языка на язык;
- Был улучшен способ совместной тренировки модели на несколько языков;

Выводы

- Мы начали с сильного baseline — BiLSTM модель с Char BiLSTM эмбеддингами;
- Был предложен более быстрый аналог Char BiLSTM, показывающий сопоставимое качество;
- Был разработан метод для предобучения эмбеддингов символьного уровня;
- Были описаны дополнительные функции потерь, улучшающие качество целевой функции — POS LM и предсказание отдельных граммем;
- Был продемонстрирован положительный эффект от переноса модели с датасета на датасет и с языка на язык;
- Был улучшен способ совместной тренировки модели на несколько языков;
- Всё это привело к значительному увеличению качества модели POS tagging'а.