

Рекуррентные нейронные сети с механизмом внимания для анализа тональности русских текстов

Иванов Илья Сергеевич¹
Ботвиновский Евгений Александрович²

¹студент, Московский Физико-Технический Институт

²к.ф.-м.н., DeepHackLab

2017

Цель исследования

Исследовать новые методы анализа тональности текстов на русском языке с применением рекуррентных нейронных сетей и механизма внимания.

Проблемы

Сложная морфология русского языка.
Особенности лексикона пользователей соц. сети.
Малый объём данных для обучения.

Предположения

Зависимость класса от порядка слов в тексте.
Разная значимость слов в тексте при классификации.

- ❶ Arkhipenko K., Kozlov I., Trofimovich J., Skorniakov K., Gomzin A., Turdakov D.. Comparison of Neural Network Architectures for Sentiment Analysis of Russian Tweets. Computational Linguistics and Intellectual Technologies. Dialog, 2016.
- ❷ Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, Eduard H. Hovy. Hierarchical Attention Networks for Document Classification. HLT-NAACL, 2016.
- ❸ Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR, 2014.

Постановка задачи классификации

Дано множество текстов (документов) $\mathcal{D} = \{\mathbf{d}_i\}_{i=1}^n$.

Необходимо классифицировать документы из \mathcal{D} на три класса:

- 1 положительной тональности (положительные);
- 2 отрицательной тональности (отрицательные);
- 3 не имеющие тональности (нейтральные).

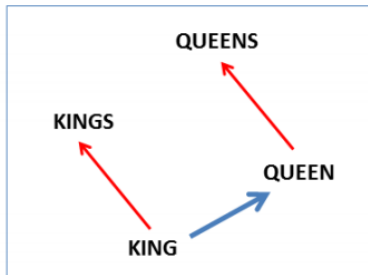
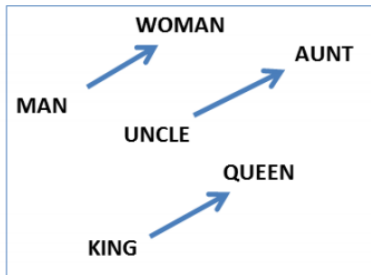
Функционалы качества

- 1 Точность (accuracy)
- 2 Макро-усредненная F-мера относительно классов положительных и отрицательных сообщений.

В качестве классификатора предлагается использовать двунаправленную рекуррентную нейронную сеть с механизмом внимания.

Векторное представление слов

- Документ $\mathbf{d} \in \mathcal{D}$ является последовательностью слов $\mathbf{d} = \mathbf{w}_1.. \mathbf{w}_T$ из словаря \mathcal{W} .
- Слово $\mathbf{w} \in \mathcal{W}$ представляется вектором в D -мерном пространстве.
- Векторное представление для всех слов из словаря получается при помощи алгоритма Word2Vec, применённом на большом наборе размеченных данных.



(Mikolov et al., NAACL HLT, 2013)

Рекуррентная нейронная сеть

- В качестве классификатора используется двунаправленная рекуррентная нейронная сеть типа GRU (Gated Recurrent Unit) с механизмом внимания.
- Функцией ошибки является перекрёстная энтропия для трёх классов.

$$J(W) = - \sum_{i=1}^n \sum_{k=1}^3 y_i^{(k)} \log \hat{y}_i^{(k)},$$

$$\hat{y}_i^{(k)} = \frac{\exp s_i^{(k)}}{\sum_{j=1}^3 \exp s_i^{(j)}}$$

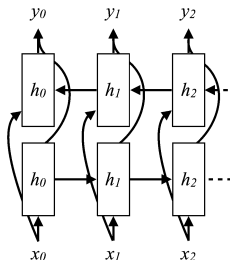
Уравнения GRU

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1}) \quad (1)$$

$$r_t = \sigma_g(W_r x_t + U_r h_{t-1}) \quad (2)$$

$$\tilde{h}_t = \tanh(W x_t + U(r_t \circ h_{t-1})) \quad (3)$$

$$h_t = (1 - z_t) \circ \tilde{h}_t + z_t \circ h_{t-1} \quad (4)$$

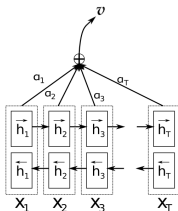


Уравнения механизма внимания

$$v_t = \tanh(W_\omega [\vec{h}_t, \overleftarrow{h}_t] + b_\omega) \quad (5)$$

$$\alpha_t = \frac{\exp(v_t^T u_\omega)}{\sum_{j=1}^T \exp(v_j^T u_\omega)} \quad (6)$$

$$v = \sum_{t=1}^T \alpha_t [\vec{h}_t, \overleftarrow{h}_t] \quad (7)$$



В качестве коллекции документов \mathcal{D} используются следующие наборы данных:

- ① Сообщения пользователей соц. сети Twitter с упоминанием некоторых банков и телекоммуникационных компаний:
 - Размер выборки - около 10 тыс. экземпляров
 - Размер сообщения - не более 140 символов
 - Лексикон: сленг, сокращения, эмодиконы
 - Спец. символы: # (хэштег), @ (ссылка на пользователя)
 - Ссылки на внешние ресурсы
- ② Отзывы на товары и рестораны:
 - Размер выборки - около 70 тыс. экземпляров
 - Размер сообщения - до 150 слов

- 1 Реализовать архитектуру двунаправленной рекуррентной сети с механизмом внимания (Python + TensorFlow)
- 2 Обучить модель на предложенных выборках
- 3 Сравнить результаты с предложенными ранее алгоритмами.

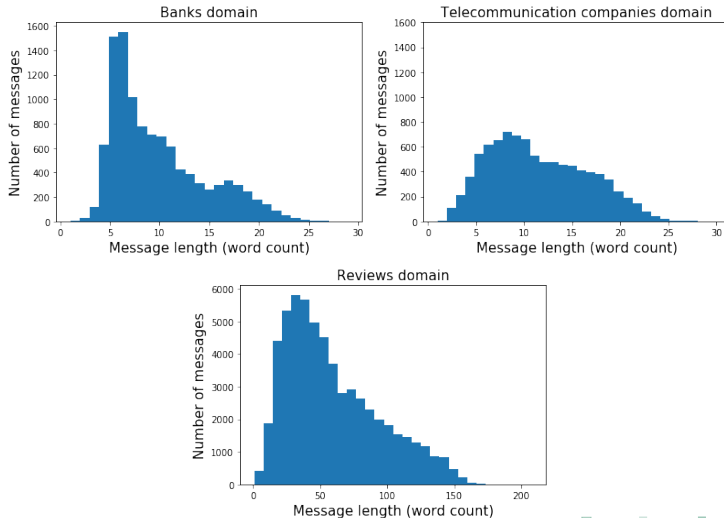
В ходе эксперимента сравниваются результаты предложенного алгоритма классификации с такими алгоритмами как двунаправленная рекуррентная нейронная сеть (без механизма внимания), метод опорных векторов и другие.

План эксперимента

- 1 Предобработать наборы текстов
- 2 Реализовать двунаправленный GRU с механизмом внимания
- 3 Провести подбор оптимальных гиперпараметров и обучить модель на обучающей выборке
- 4 Протестировать модель на отложенной выборке
- 5 Сравнить результаты с другими алгоритмами

- 1 Токенизация (NLTK)
- 2 Лемматизация (PyMorphy2)
- 3 Векторизация слов (Word2Vec, обученный на русскоязычном корпусе из социальных медиа)
- 4 Дополнение последовательностей нулями до максимальной длины (zero-padding)

Рис.: Распределение кол-ва слов в сообщении



Визуализация механизма внимания

1. Почему-то не приходят смс-сообщения для подтверждения входа
2. iPhone-овское приложение от Сбербанка самое удобное в России
3. Заведение вполне приличное, кухня хорошая, но маловато выбора, зато с напитками никакой проблемы выбора нет!! много сортов пива и других более крепких напитков. из минусов можно сказать только чрезмерная громкость живой музыки по выходным. соседа не слышно....

Сравнение качества полученных моделей

Таблица: F1-мера различных моделей на кросс-валидации (CV) и на тестовой выборке для выборок с твитами

	Banks		Telecommunication companies	
	5-fold CV (mean, std)	test	5-fold CV (mean, std)	test
Bi-GRU	0.74, 0.02	0.48	0.62, 0.01	0.52
Bi-GRU + Attention	0.74, 0.02	0.51	0.60, 0.02	0.49
2-layer GRU, reversed sequences (Arhipenko)	0.62, -	0.55	0.66, -	0.56
Bi-GRU (Arhipenko)	0.62, -	-	0.65, -	-
LSTM (Arhipenko)	0.60, -	-	0.64, -	-
CNN (Arhipenko)	-	0.48	-	0.47
SVM baseline	-	0.46	-	0.46
Majority baseline	-	0.31	-	0.19

Сравнение качества полученных моделей

Таблица: Результаты эксперимента со смешиванием обучающей и тестовой выборок

	Banks			Telecommunication companies		
	cross-validation		test	cross-validation		test
	train	train+test		train	train+test	
Bi-GRU	0.74, 0.02	0.71, 0.02	0.48	0.62, 0.01	0.62, 0.01	0.52
Bi-GRU+Attention	0.74, 0.02	0.72, 0.01	0.51	0.60, 0.02	0.62, 0.01	0.49

Таблица: Качество различных моделей на кросс-валидации (CV) и на тестовой выборке для выборки с отзывами

	Reviews			
	10-fold CV (mean, std)		test	
	Accuracy	F1	Accuracy	F1
Bi-GRU	0.906, 0.003	0.863, 0.007	0.901	0.861
Bi-GRU+Attention	0.907, 0.004	0.865, 0.007	0.900	0.861

- Реализован алгоритм двунаправленной рекуррентной нейронной сети с механизмом внимания для классификации тональности русскоязычных текстов. Код отлажен и выложен в открытый доступ
- Проведён подбор гиперпараметров и обучены модели на вышеупомянутых выборках
- Проведено сравнение результатов с предложенными ранее алгоритмами
- Подготовлен отчет по результатам работы

Дальнейшее исследование

Проведение экспериментов на других наборах данных. Исследование применимости данной модели в качестве модуля для нейронной сети, генерирующей сообщения с заданной тональностью.