УДК 004.8

А. А. Харламов[1], *Ле Мань Ха*[2]

[1]Институт высшей нервной деятельности и нейрофизиологии РАН
[2]Московский физико-технический институт (государственный университет)

Нейросетевые подходы к классификации текстов на основе морфологического анализа

В данной работе было рассмотрено применение морфологического анализа в классификации текстов. Морфологический анализ позволяет изучать грамматические свойства слов, а также грамматическую семантику и взаимодействие между элементами текстов. Были предложены нейро-семантическая сеть на основе морфологического анализа для изучения векторных представлений грамматических структур текста и рекурсивный автоэнкодер, который состоит из двух частей: первая объединяет два вектора слов, а вторая – два вектора морфологий.

Ключевые слова: классификация, морфология, нейронная сеть

A. A. Kharlamov[1], *Le Manh Ha*[2]

[1]Institute of Higher Nervous Activity and Neurophysiology of RAS
[2]Moscow Institute of Physics and Technology (State University)

Neural network approaches to the classification of texts based on morphological analysis

In this paper, the main purpose is to consider applications of morphological analysis in text classification. Morphological analysis helps us to learn grammatical features of words, grammatical semantic and the interaction between the elements of text. We propose the neurosemantic network based on morphological analysis for learning vector representations of the text's grammatical structures and the recursive autoencoder that consists of two parts - the first part combines two vectors of words, the second one combines two vectors of morphology.

Key words: classification, morphology, neural network

1. Введение

За последние несколько лет нейронные сети вновь появились в качестве мощных моделей машинного обучения, показали лучшие результаты в таких областях, как распознавание образов и обработки речи. Еще совсем недавно нейросетевые модели начали применяться также к различным задачам обработки естественного языка с очень хорошими результатами [1]. Традиционная модель «мешок слов» вместе с классификаторами, которые используют эту модель, такими, как байесовский метод, были успешно использованы для того, чтобы получать очень точные прогнозы в задаче анализа настроений [2]. С появлением технологий глубокого обучения и их применения в обработке естественного языка было сделано улучшение точности этих методов в двух основных направлениях: использование нейронной сети с учителем для обучения классификатора и без учителя для оптимизации предварительной обработки данных и выбора характеристик.

Морфология имеет важное значение в обработке естественного языка, в том числе для анализа и генерации текстов. Существует широкий спектр таких приложений, например, извлечение информации из текстовых сообщений или диалоговых систем, машинный перевод [3]. Морфологический анализ является важной частью синтаксического анализа. Свойства слов, полученные в результате морфологического анализа (леммы и его части речи), используются для глубокого анализа текстов. Другой важной задачей является лематизация – поиск соответствующей словоформы для данного входного слова.

2. Нейросетевые подходы к классификации текстов на основе морфологического анализа

2.1. Морфологический анализ

Морфологический анализ – процесс поиска морфологических разборов слов. Цель морфологического анализа – выяснить, из каких морфем построены слова. Например, морфологический анализатор должен сказать, что слово «кошки» является формой множественного числа существительного «кошка», и слово «мыши» является формой множественного числа существительного «мышь». Таким образом, принимая слово «кошек» в качестве входных данных, морфологический анализатор должен производить выход, похожий на «кошка NOUN femn plur gent».

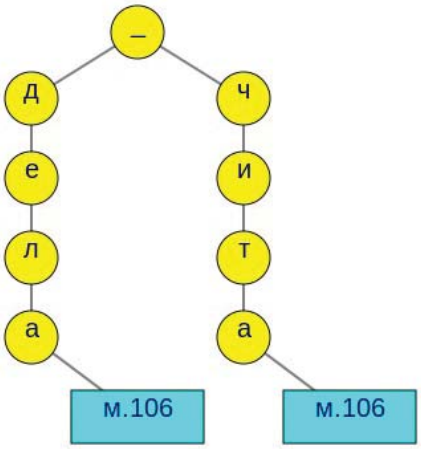

Рис. 1. Структура морфологического словаря

Для реализации морфологического словаря используется структура данных «префиксное дерево» [4] – это тип дерева поиска для хранения ассоциативного массива из элементов (ключ, значение), где ключи являются префиксами строк. Ключ одного узла состоит из символов на пути из корня дерева до этого узла. Корень дерева содержит пустую строку. Значения, связанные с ключом, содержат морфологические модели префикса этого узла. Чтобы найти морфологические варианты словоформы, нужно обходить дерево по символам этой словоформы. Временная сложность операции морфологического поиска словоформы является линейной и равна $O(n)$, где n – длина словоформы.

Пример модели русской морфологии № 106:

Модель	Суффикс	Разметка
106	-ть	impf, tran, VERB
	-ю	1per, indc, pres, sing
	-ем	1per, indc, plur, pres
	-ешь	2per, indc, pres, sing
	-ете	2per, indc, plur, pres
	-ет	3per, indc, pres, sing
	-ют	3per, indc, plur, pres
	-л	indc, masc, past, sing
	-ла	femn, indc, past, sing
	-ло	indc, neut, past, sing
	-ли	indc, past, plur
	-й	excl, impr, sing
	-йте	excl, impr, plur

2.2. Нейро-семантическая сеть на основе морфологического анализа – Morphological Neural Semantic Networks – MNSN

Нейро-семантическая сеть состоит из трех последовательных частей.

1. Часть «семантические векторные представления», которая вычисляет векторные представления грамматических структур предложений, содержит автоэнкодеры по заданным грамматическим структурам (SVO, SVA, ...), которые принимают на вход слова в виде пар (вектор, морфология) и объединяют их в одну пару. Цель заключается в том, чтобы близкие по смыслу структуры предложений имели похожие векторные представления, например: «девушка читает книгу» и «женщина читает роман».

Примеры часто встречаемых грамматических структур:

- AN (прилагательное – существительное);
- SVO (субъект–действие–объект);
- SVA (субъект–действие–наречие).

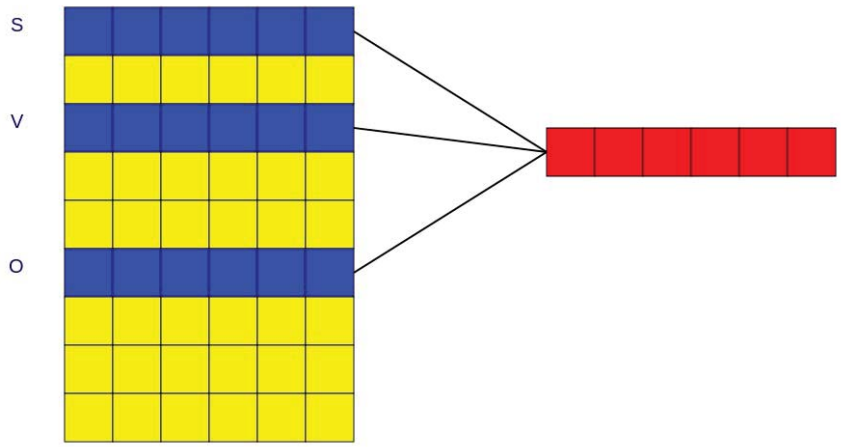

Рис. 2. Семантическое векторное представление

Для поиска грамматических структур для обучения семантических векторных представлений будем считать частоту вхождения множеств из не более чем n_g семантических моделей. Задан семантический разбор текста в виде последовательности из n семантических моделей $m_1, m_2, ..., m_n$, рассмотрим:

Для $i_1 = 1..n$:
  Для $i_2 = i_1 + 1..n$:
  ...
    Для $i_k = i_{k-1} + 1..n (k \leqslant n_g)$:
      Сортировать $m_{i_1}, m_{i_2}, ..., m_{i_k}$ по возрастанию, получаем $\hat{m}_{i_1}, \hat{m}_{i_2}, ..., \hat{m}_{i_k}$.
      Обновить частоту для множеств $\hat{m}_{i_1}, \hat{m}_{i_2}, ..., \hat{m}_{i_k}$.

В результате получаем набор самых встречаемых множеств семантических моделей.

Для обучения семантических векторных представлений используются автоэнкодеры [5] размера $n_g \times k_{word}$, где n_g – количество элементов в грамматической структуре (например, $n_g = 2$ для AN, $n_g = 3$ для SVO и SVA), а k_{word} – размерность векторного пространства слов. Для каждой грамматической структуры используется отдельный автоэнкодер для обучения представлений. Входной и выходной слои автоэнкодера имеют $n_g \times k$ нейронов, а скрытый слой – $k_{sementic}$ нейронов ($k_{sementic}$ – размерность пространства семантических векторных представлений):

Автоэнкодер объединяет элементы грамматической структуры $x_1, x_2, ..., x_{n_g}$ (n_g-вектора длины k_{word}) в один вектор y (длины $k_{sementic}$):

$$y = f(W^{(1)}[x_1, x_2, ..., x_{n_g}] + b^{(1)}). \tag{1}$$

Чтобы вычислить ошибку объединения, нужно восстанавливать исходные слова $x_1, x_2, ..., x_{n_g}$ из вектора y:

$$[\acute{x_1}, \acute{x_2}, ..., \acute{x_{n_g}}] = W^{(2)}y + b^{(2)}. \tag{2}$$

Ошибка объединения есть евклидово расстояние между парами исходных и восстановленных слов:

$$E_{rec}([x_1, x_2, ..., x_{n_g}]) = \|x_1 - \acute{x_1}\|^2 + \|x_2 - \acute{x_2}\|^2 + ... + \|x_{n_g} - \acute{x_{n_g}}\|^2. \tag{3}$$

Чтобы ошибка объединения не уменьшилась просто из-за уменьшения скрытых слоев, используется нормализация объединенного вектора:

$$\hat{y} = \frac{y}{\|y\|}. \tag{4}$$

2. Часть «распределение по категориям семантического представления», которая принимает на вход объединенный вектор x и вычисляет распределение категорий h по предложению, является слоем Softmax [6].

3. Часть «распределение по категориям текста», которая принимает на вход распределения категорий по предложениям и вычисляет распределение категорий по тексту. Распределение вероятностей для текста, состоящего из N предложений, есть среднее распределение вероятностей по предложениям:

$$h_\theta = \frac{1}{N}\sum_{t=1}^{N} h_\theta^{(t)}. \tag{5}$$

Нейро-семантическая сеть на основе морфологического анализа имеет преимущество перед обычным рекурсивным автоэнкодером в том, что она учитывает грамматическую структуру текстов, поэтому векторные представления грамматических структур отражают семантические значения текстов.

Целью здесь является создание семантических векторных пространств и поиск функций, которые вычисляют семантические векторные представления грамматических структур по векторным представлениям слов, такие как:

$$v_{AN}(red \quad car) = f_{AN}(v(red), v(car)), \tag{6}$$

$$v_{AN}(blue \quad car) = f_{AN}(v(blue), v(car)), \tag{7}$$

$$v_{SVO}(\text{я читаю книгу}) = f_{SVO}(v(\text{я}), v(\text{читать}), v(\text{книга})), \tag{8}$$

где $v_{AN}(red \quad car), v_{SVO}(\text{я читаю книгу})$ – семантические векторные представления для «red car» и «я читаю книгу», $v(red), v(blue), v(car), v(\text{я}), v(\text{читать}), v(\text{книга}$ – векторные представления, f_{AN} и f_{SVO} – искомые функции.

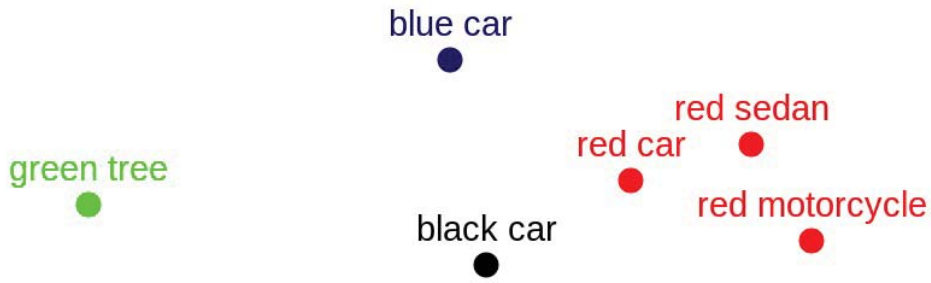

Рис. 3. Семантическое векторное пространство

2.3. Рекурсивный автоэнкодер морфологического анализа – Morphological Recursive AutoEncoder – MRAE

Предлагается рекурсивный автоэнкодер [7] морфологического анализа, который состоит из двух частей: первая объединяет два вектора слов, а вторая объединяет два вектора морфологий. Морфологическая часть рекурсивного автоэнкодера позволяет повышать точность выбора векторов слов в процедуре формирования векторного представления текста.

Векторное представление текста получается повторением процесса объединения двух слов-векторов с использованием рекурсивного автоэнкодера. На каждом этапе выбор пары слов-векторов объединение происходит с помощью данного рекурсивного автоэнкодера.

Автоэнкодер объединяет две пары (слово; морфологический разбор) $(x_1; m_1), (x_2, m_2)$ в одну пару (y, m):

$$y = f(W_w^{(1)}[x_1, x_2] + b_w^{(1)}), \tag{9}$$

$$m = f(W_m^{(1)}[m_1, m_2] + b_m^{(1)}). \tag{10}$$

Чтобы вычислять ошибки объединения, нужно восстанавливать исходные слова $x1, x2$ из вектора y:

$$[\acute{x_1}, \acute{x_2}] = W_w^{(2)} y + b_w^{(2)}, \tag{11}$$

$$[\acute{m_1}, \acute{m_2}] = W_m^{(2)} m + b_m^{(2)}. \tag{12}$$

Ошибки объединения есть евклидово расстояние между парами исходных и восстановленных слов:

$$E_w([x_1, x_2]) = \frac{n_1}{n_1 + n_2}\|x_1 - \acute{x_1}\|^2 + \frac{n_2}{n_1 + n_2}\|x_2 - \acute{x_2}\|^2, \tag{13}$$

$$E_m([m_1, m_2]) = \frac{n_1}{n_1 + n_2}\|m_1 - \acute{m_1}\|^2 + \frac{n_2}{n_1 + n_2}\|m_2 - \acute{m_2}\|^2, \tag{14}$$

$$E_{rec} = \alpha E_w + (1 - \alpha) E_m, \tag{15}$$

где n_1, n_2 – количества объединенных слов в x_1, x_2 соответственно; α – взвешенный параметр для ошибок объединений слов и морфологических разборов.

Чтобы ошибка объединения не просто уменьшалась из-за уменьшения значений нейронов скрытых слоев, будем нормализовать объединенные вектора слов и морфологии:

$$\hat{y} = \frac{y}{\|y\|}, \tag{16}$$

$$\hat{m} = \frac{m}{\|m\|}. \tag{17}$$

Этот процесс повторяется $N-1$ раз для текста, состоящего из N слов. В результате получается окончательный вектор – семантическое векторное представление текста, этот вектор используется как вход для системы обучения.

Распределение вероятностей для K-классовой классификации для векторного представления текста y вычисляется слоем Softmax [6]:

$$d(y;\theta) = Softmax(W^{Softmax}y). \tag{18}$$

Ошибка регрессии для K-классовой классификации записывается как перекрёстная энтропия между выходным и целевым распределениями:

$$E_{reg} = -\sum_{k=1}^{K} t_k \log d_k. \tag{19}$$

Ошибка объединения на узле j:

$$E_{mer_j} = \beta E_{rec} + (1-\beta)E_{reg}, \tag{20}$$

где β – взвешенный параметр для ошибок регрессии и объединения.

Суммарная ошибка на всех узлах текста i:

$$E_i = \sum_{j=1}^{N-1} E_{mer_j}. \tag{21}$$

Функция потерь автоэнкодера есть средняя ошибка для всех текстов в обучаемой выборке из M текстов:

$$J = \frac{1}{M}\sum_{i=1}^{M} E_i + \frac{\lambda}{2}\|\theta\|^2. \tag{22}$$

Градиент функции потерь:

$$\frac{\partial J}{\partial \theta} = \frac{1}{M}\sum_{i=1}^{M} \frac{\partial E_i}{\partial \theta} + \lambda\theta. \tag{23}$$

3. Эксперименты и оценка результатов реализуемых алгоритмов

Для оценки алгоритмов был использован метод K-Fold Cross-validation [8] с параметром $K = 10$. В качестве экспериментальной базы использованы базы Movie Review и Wikinews на русском и английском языках. Обучающая выборка Movie Review состоит из 10662 текстов по двум категориям (позитивной и негативной). Обучающая выборка Wikinews-Ru состоит из 7233 текстов – новостей на русском языке, её количество категорий равно 8. Обучающая выборка Wikinews-En состоит из 23588 текстов – новостей на английском языке, её количество категорий равно 11.

Классификация базы данных Movie Review:

Метод	Movie Review
RAE (Socher et al., 2011)	77.7
MV-RNN (Socher et al., 2012)	79.0
CNN-rand	76.1
CNN-static	81.0
CNN-non-static	81.5
CNN-multichannel	81.1
MNSN	79.3
MRAE	80.6

Классификация баз данных Wikinews:

Метод	Wikinews-Ru	Wikinews-En
Naive Bayes	66.4	81.1
Nearest Centroid	68.6	86.3
KNN	72.5	87.3
Оптимизированный KNN	71.7	88.9
MNSN	75.2	91.3
MRAE	74.3	90.2

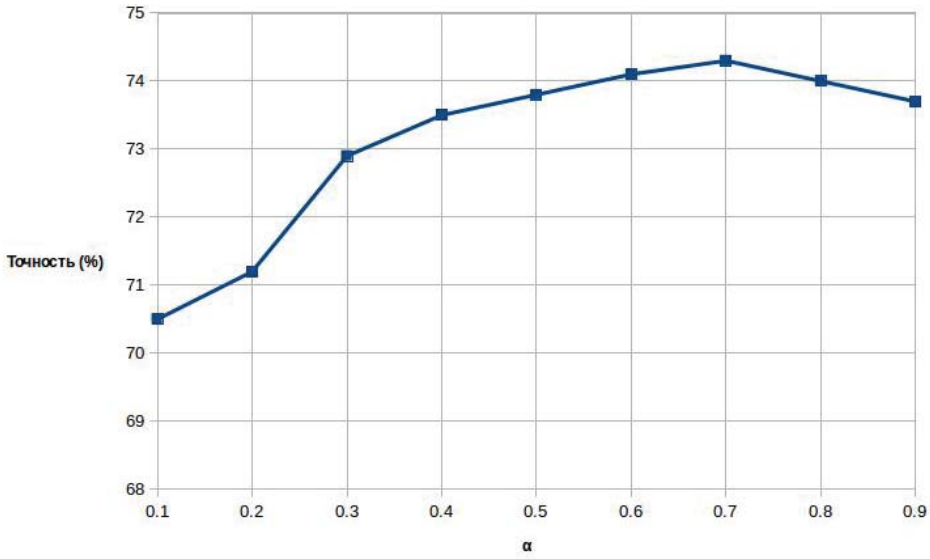

Рис. 4. Точность классификации базы Wikinews-En рекурсивным автоэнкодером морфологического анализа в зависимости от взвешенного параметра α для ошибок объединений слов и морфологических разборов

4. Заключение

Учитывая результаты экспериментов, можно прийти к выводу, что морфологический анализ позволяет изучать морфологические свойства слов, а также грамматическую се-

мантику и взаимодействие между элементами текстов, что позволяет изучать смысловое значение и структуру текстов и повышать качество классификации текстов.

Литература

1. *Lee J.Y., Dernoncourt F.* Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks // Proceedings of NAACL-HLT. 2016. P. 515–520.
2. *Narayanan V., Arora I., Bhatia A.* Fast and accurate sentiment classification using an enhanced Naive Bayes model // International Conference on Intelligent Data Engineering and Automated Learning. 2013. Oct 20. P. 194–201. Springer Berlin Heidelberg.
3. *Carstairs-McCarthy A.* An Introduction to English Morphology: Words and Their Structure. Edinburgh: Edinburgh University Press, 2002.
4. *Knuth D.E.* The Art of Computer Programming: Volume 3: Sorting and Searching. Addison-Wesley Professional, 1998. Apr 24.
5. *Hinton G.E., Salakhutdinov R.R.* Reducing the dimensionality of data with neural networks. science. 2006. Jul 28. 313(5786):504-7.
6. *Memisevic R., Zach C., Pollefeys M., Hinton G.E.* Gated softmax classification // Advances in neural information processing systems. 2010. P. 1603–1611.
7. Socher R, Pennington J, Huang EH, Ng AY, Manning CD. Semi-supervised recursive autoencoders for predicting sentiment distributions // Proceedings of the conference on empirical methods in natural language processing 2011 Jul 27. Association for Computational Linguistics. P. 151–161.
8. *Efron B., Tibshirani R.* Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science. 1986. Feb 1:54-75.

References

1. *Lee J.Y., Dernoncourt F.* Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. Proceedings of NAACL-HLT. 2016. P. 515–520.
2. *Narayanan V., Arora I., Bhatia A.* Fast and accurate sentiment classification using an enhanced Naive Bayes model. International Conference on Intelligent Data Engineering and Automated Learning. 2013. Oct 20. P. 194–201. Springer Berlin Heidelberg.
3. *Carstairs-McCarthy A.* An Introduction to English Morphology: Words and Their Structure. Edinburgh: Edinburgh University Press, 2002.
4. *Knuth D.E.* The Art of Computer Programming: Volume 3: Sorting and Searching. Addison-Wesley Professional, 1998. Apr 24.
5. *Hinton G.E., Salakhutdinov R.R.* Reducing the dimensionality of data with neural networks. science. 2006. Jul 28. 313(5786):504-7.
6. *Memisevic R., Zach C., Pollefeys M., Hinton G.E.* Gated softmax classification. Advances in neural information processing systems. 2010. P. 1603–1611.
7. Socher R, Pennington J, Huang EH, Ng AY, Manning CD. Semi-supervised recursive autoencoders for predicting sentiment distributions. Proceedings of the conference on empirical methods in natural language processing 2011 Jul 27. Association for Computational Linguistics. P. 151–161.
8. *Efron B., Tibshirani R.* Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science. 1986. Feb 1:54-75.

Поступила в редакцию 31.05.2017