



Ю. П. Сердюк, Н. А. Власова

Распознавание идиоматического использования выражений с помощью нейронных сетей

Аннотация. Многие идиоматические выражения могут использоваться не только в переносном смысле, но и в прямом. Распознавание того или иного случая их употребления является важной задачей во многих приложениях обработки текстов на естественном языке, в частности, в машинном переводе. В настоящей работе предлагается автоматический способ распознавания прямого и переносного использования идиоматических выражений на основе анализа их локальных контекстов с помощью рекуррентных нейронных сетей. Исследованы два типа таких сетей для решения данной задачи — обычные рекуррентные нейросети и двунаправленные их модификации. Рассмотрены варианты представления слов контекста как в виде нормальных форм, так и в виде словоформ, встретившихся в тексте. Описаны способ построения и характеристики дистрибутивной модели, в которой хранятся векторные представления слов и целевых идиоматических выражений. В заключение мы даем обзор наиболее важных работ по данной проблематике.

Ключевые слова и фразы: идиоматические выражения, нейронные сети, рекуррентные нейронные сети, векторные представления слов и выражений, распознавание именованных сущностей.

Введение

Интерпретация образных (идиоматических, метафорических) выражений является одной из трудных задач в области автоматической обработки текстов на естественном языке. В частности, некоторые идиоматические выражения могут употребляться не только в своем переносном (некомпозициональном) значении, которое требует их

Работа выполнена при поддержке РФФИ (проект № 19-07-00991-а).

© Ю. П. Сердюк, Н. А. Власова, 2021

© Институт программных систем им. А. К. Айламазяна РАН, 2021

© Программные системы: теория и приложения (дизайн), 2021

 10.25209/2079-3316-2021-12-3-3-27



специальной трактовки, но и в прямом смысле. Такие примеры есть и в русском языке:

Пример 1. Мужчина с огнестрельными ранениями **правой руки** был доставлен в центральную городскую больницу.

В настоящее время Делимханов считается **правой рукой** и возможным преемником главы Чечни Рамзана Кадырова.

Так, например, изучение 60 английских идиом показало, что около половины из них используется и в прямом значении [1]. Более того, для некоторых идиоматических выражений такое употребление может преобладать [2]. В русском языке многословные сочетания также могут употребляться как в прямом, так и в переносном значении.

Некоторые из таких многословных выражений могут употребляться как метафоры. В настоящей работе метафорические употребления многословных выражений мы будем относить к переносному значению¹.

Пример 2. Его **длинный язык** сыграет с ним злую шутку. (*переносное значение, идиома*).

У жирафа очень гибкий **длинный язык**. (*прямое значение*).

К небу взвивались **длинные языки** пламени. (*переносное значение, метафора*).

В данной статье предлагается автоматический способ распознавания прямого/переносного использования идиоматических выражений на основе анализа их локальных контекстов. Во многих случаях такие контексты довольно точно определяют значение многословных сочетаний и поэтому могут служить входом для нейронной сети, осуществляющей классификацию таких выражений по их прямому или переносному использованию. Способ представления контекстов идиом для эксперимента в настоящем исследовании описан в разделе 1.

Анализ контекстов различного размера (от нескольких слов слева и справа от целевого выражения до предложения целиком) является базовым подходом во всех работах, относящихся к данной тематике. Отличия заключаются в способах представления таких контекстов. При применении нейронных сетей в качестве классификатора способов употребления идиоматических выражений важное значение имеют

¹Обзор методов классификации выражений по принципу «метафора — не метафора» дан в диссертации [3].

векторные представления входных слов и выражений. Впервые такие представления были использованы для решения данной задачи в работе [4]. Однако, если в ранних работах векторные представления получались с помощью сингулярного разложения (Singular Value Decomposition, SVD) матриц совместного вхождения слов, как в методе латентного семантического анализа (Latent Semantic Analysis), или ковариационной матрицы, как в методе главных компонент (Principal Component Analysis, PCA), то в большинстве последних работ исследователи опираются на векторные представления, которые получены на основе широко распространенного стандартного метода word2vec.

Как показано в настоящей работе, использование векторных представлений слов контекста и целевых идиоматических выражений, полученных word2vec-методом на большом корпусе текстов (около 160 млн. предложений), позволяет достичь довольно хороших показателей точности распознавания значения идиом — от 86% до 96% в зависимости от конкретного выражения. Обзор наиболее важных работ по данной проблематике, использующих различные виды контекстов и способы их классификации, представлен в разделе 4.

В данном исследовании для проведения экспериментов было выбрано 10 широко распространенных идиоматических выражений русского языка, используемых также в прямом значении: *правая рука, темный лес, серая мышь, вчерашний день, тяжелый груз, длинный язык, грязные руки, красная линия, закрытая дверь, последнее слово*. Все эти идиоматические выражения имеют одинаковую синтаксическую структуру, а именно — имя прилагательное + имя существительное.

Был подготовлен корпус предложений, в которых встречаются вышеперечисленные многословные сочетания — не менее 600 предложений (300 + 300 для прямого и переносного значений) для каждого целевого выражения, что дало, в конечном итоге, корпус объемом более 6000 предложений, из них 4000 предложений были использованы в качестве обучающего множества, а оставшиеся 2000 — в качестве тестового.

Отметим, что для русского языка единственным известным корпусом для решения аналогичной задачи является аннотированный корпус русских идиом из 5,5 тыс. примеров их прямого/переносного употреблений [5], в нем представлены идиомы различной синтаксической структуры, в частности с предлогами (*на свою голову,*

смотреть в глаза и др.), тогда как для создания нашего корпуса подбирались только многословные сочетания вида имя прилагательное + имя существительное. Наш эксперимент и его результаты описаны в разделе 2².

Несмотря на одинаковую синтаксическую структуру, каждая из исследованных в настоящей статье идиом имеет свои специфические семантические и другие свойства, с учетом которых можно более качественно решать задачу классификации употребления выражений в прямом или переносном смысле. Эти свойства для 10 обсуждаемых в работе идиоматических выражений представлены в разделе 3.

Также в этом разделе описаны решения, которые были реализованы в предлагаемом подходе, — выявление и факторизация собственных имен и некоторых других именованных сущностей (названий организаций, географических названий), создание векторных представлений для знаков препинания и использование таких представлений в локальных контекстах идиом и др.

В заключительном разделе 5 представлены идеи, реализация которых позволит, по нашему мнению, решить трудные примеры классификации прямого/переносного использования идиоматических выражений, с которыми описанный в данной работе подход пока не справляется.

1. Принципы классификации прямого/переносного использования идиоматических выражений

Главный принцип нашего подхода к распознаванию идиоматического употребления многословных выражений состоит в том, что его вид можно определить, используя:

- (1) (локальный) контекст целевого выражения;

²В работе [6] описан корпус из 220 двухсловных выражений со структурой «прилагательное-существительное» и «существительное-существительное», аннотированных как «композиционные», «некомпозиционные» и «неоднозначные». В последнем случае степень композиционности выражения зависит от контекста его употребления. Однако данный корпус не содержит аннотированных по указанному принципу примеров (предложений) использования таких выражений. Соответственно, распознавание типов употребления выражений на основе контекстов отнесено авторами в качестве будущей работы.

- (2) форму, в которой употреблены слова этого выражения в заданном предложении;
- (3) знаки препинания;
- (4) порядок слов в предложении.

Локальный контекст в нашем случае состоит из двух лексических единиц, взятых слева и справа от целевого выражения. Если в зависимости от структуры предложения отдельные слова контекста отсутствуют (из-за границ предложения), то вместо них подставляется специальный элемент **<out-of-bounds>**. Размер окна (лексического контекста), равный двум, является стандартным в различных задачах NLP. С другой стороны, этот параметр оказывается специфичным для разных идиом — так, например, идиома *грязные руки* часто располагается в конце предложения, а потому более, чем в 25% обучающих примеров для нее уже (при размере окна 2) присутствуют элементы типа **<out-of-bounds>**. Исходя из этих соображений и был выбран единый для всех идиом размер локального контекста.

Соответственно, элементы обучающего и тестового множеств, которые выделялись из предложений корпуса, имеют следующую общую структуру:

W_{-2}	W_{-1}	идиоматическое выражение	W_{+1}	W_{+2}
----------	----------	-----------------------------	----------	----------

Рисунок 1. Структура обучающего/тестового примера

Использование идиоматического выражения в том виде, в котором оно присутствует в заданном предложении (а не в канонической форме), в качестве элемента обучающего/тестового множества обусловлено тем фактом, что сами формы слов, составляющих многословное выражение, могут говорить о способе его интерпретации.

Пример 3. А женские причуды для меня вообще **темный лес** ...

Унесут за **темные леса**, за высокие горы ...

Для многословного выражения *темный лес* использование его во множественном числе почти всегда говорит о прямом его применении. Аналогичная ситуация наблюдается и для многих других идиоматических выражений со структурой «*прилагательное+существительное*».

Порядок слов в предложении (а значит, и в локальном контексте) учитывается в нашем подходе путем применения для распознавания рекуррентных нейронных сетей (в том числе двунаправленных), которые считывают и обрабатывают слова входной последовательности (обучающего/тестового примера) друг за другом.

Аналогично знаки препинания также часто добавляют информации о том, в каком смысле — прямом или переносном — используется идиоматическое выражение.

Пример 4. В этот момент дверь хижины отворилась и внутрь проскочила маленькая **серая мышь**.

Ты такая красивая и интересная, а я невзрачная **серая мышь**!

Данный пример иллюстрирует ситуацию, в которой для идиоматического выражения «*серая мышь*» использование его совместно с восклицательным знаком чаще всего говорит о переносном применении этого выражения.

Обычной ситуацией при использовании дистрибутивных моделей языка является отсутствие в словаре модели некоторых слов, встречающихся в предложениях, на основе которых строились примеры обучающего и тестового множеств. Чаще всего такими словами являются редкие собственные имена, географические названия и другие именованные сущности.

Чтобы избежать соответствующих «пробелов» в обучающих/тестовых примерах, мы осуществили предобработку предложений с целью обнаружения в них именованных сущностей (**named entities**) трех типов:

- (1) собственных имен людей (сущностей типа **PERSON**);
- (2) географических названий (сущностей типа **LOCATION**);
- (3) названий организаций (сущностей типа **ORGANIZATION**).

Для этого мы использовали систему распознавания именованных сущностей из пакета **deeppavlov** [7], в которой, в свою очередь, мы применили модель **ner_rus_bert**³. Обнаруженные именованные сущности (в общем случае, многословные) заменялись в предложениях (и, соответственно, в обучающих/тестовых примерах) на имена классов **PER**, **LOC** и **ORG**, соответственно. Если некоторое слово, тем не менее,

³<http://docs.deeppavlov.ai/en/master/features/models/ner.html>

отсутствует в словаре дистрибутивной модели и не входит в состав именованных сущностей типа PER, LOC и ORG, то такому слову в обучающем/тестовом примере сопоставлялся специальный символ (знак нижнего подчеркивания), и специальный «нулевой» вектор в векторном представлении слов. Ниже показано несколько характерных примеров из обучающих/тестовых множеств для различных целевых идиом.

Пример 5.

[«встреча»	«за»	«закрытая дверь»	«еще»	«продолжаться»]
[«к»	PER	«грязные руки»	«.»	«out-of-bounds»]
[LOC	«и»	«темный лес»	«под»	«гора»]

[«встреча»	«за»	«закрытыми дверями»	«еще»	«продолжается»]
[«к»	PER	«грязную руку»	«.»	«out-of-bounds»]
[LOC	«и»	«темный лес»	«под»	«горой»]

Первые три примера, из показанных выше, построены на основе стандартных форм, а следующие три — эти же три примера, но уже на основе исходных (произвольных) словоформ.

В наших экспериментах мы сравнили два варианта распознавания прямого/переносного использования идиоматических выражений (см. об этом более подробно в разделе 2.2):

- (1) на основе использования стандартных форм (лемм) отдельных слов и целевых выражений (соответственно, дистрибутивной модели, содержащей только леммы слов и стандартные формы выражений);
- (2) на основе произвольных словоформ (соответственно, «большой» дистрибутивной модели, содержащей все возможные словоформы отдельных слов и целевых идиоматических выражений).

2. Результаты эксперимента

2.1. Корпус идиоматических выражений

Для проведения экспериментов по распознаванию, было выбрано 10 идиоматических выражений со структурой «*прилагательное+существительное*», употребляемых как в прямом, так и переносном смыслах:

- | | |
|--------------------------|----------------------------|
| 1) <i>правая рука</i> | 6) <i>длинный язык</i> |
| 2) <i>темный лес</i> | 7) <i>грязные руки</i> |
| 3) <i>серая мышь</i> | 8) <i>красная линия</i> |
| 4) <i>вчерашний день</i> | 9) <i>закрытая дверь</i> |
| 5) <i>тяжелый груз</i> | 10) <i>последнее слово</i> |

Для каждого из этих выражений было выбрано 400 предложений для построения обучающего множества (200 в прямом смысле и 200 в переносном смысле) и 200 предложений для тестового множества (100+100 прямой/переносный смыслы, соответственно). Предложения выбирались из трех источников:

- (1) корпуса PaRuS⁴;
- (2) корпуса Taiga⁵;
- (3) из Интернета с помощью Google Search API.

Основная часть предложений была взята из корпуса PaRuS [8], общий объем которого составляет около 160 млн. предложений. Тем не менее, для отдельных употреблений некоторых идиоматических выражений в этом корпусе набралось недостаточно предложений (в частности, предложений с прямым использованием выражения *серая мышь*), и потому до нужного количества предложения добирались из других источников. Более того, поскольку мы в качестве одного из исследуемых вариантов использовали дистрибутивную модель с произвольными словоформами идиоматических выражений, для некоторых таких словоформ оказалось слишком мало предложений, чтобы можно было построить качественную дистрибутивную модель. Таким образом, этот опыт показывает, что для полноценной работы со всеми возможными видами словоформ идиоматических выражений необходимы корпуса с не менее, чем 1 млрд. предложений.

2.2. Дистрибутивная модель слов и выражений

Включение идиоматического выражения в состав обучающего/тестового примера требует в дальнейшем применения расширенной дистрибутивной модели. Словарь такой модели состоит из одиночных слов (как в стандартных дистрибутивных моделях) и идиоматических

⁴<https://parus-proj.github.io/PaRuS/>

⁵https://tatianashavrina.github.io/taiga_site/

выражений целиком. Для наших экспериментов мы построили два варианта таких расширенных моделей:

- (1) на основе канонических форм (лемм) слов и целевых идиоматических выражений;
- (2) на основе их произвольных форм.

Словарь первой модели составил 112 300 слов, длина вектора была выбрана равной 100. Размер словаря второй модели — 1 740 000 слов, длина вектора — 300. Исходным корпусом для построения обеих моделей был корпус PaRuS. Для построения первой модели, стандартные формы слов брались из самого корпуса. Для включения идиоматических выражений в дистрибутивные модели, весь корпус был предобработан путем замены этих выражений на отдельные лексические единицы — токены (т.е., например, идиома *красная линия* заменялась на токен *красная_линия*).

Для подготовки векторной модели, предложения из обучающего и тестового множеств предобрабатывались с помощью подсистемы распознавания именованных сущностей (NER) на базе системы *deeppavlov*. F1-показатель этой подсистемы заявлен её авторами в 98.1%. Распознанные именованные сущности трех классов заменялись на соответствующие имена классов PER, LOC и ORG. В свою очередь, эти имена классов при построении векторного представления предложений заменялись на вектора специального вида (конкретно, это были «one-hot»-вектора, т.е., вектора с единственной единицей в некоторой позиции и с нулями во всех остальных). Аналогичными векторами кодировались слова, отсутствующие в словаре модели, а также элементы обучающего/тестового множеств вида <out-of-bounds>. Поскольку количество предложений с некоторыми словоформами отдельных идиоматических выражений даже в корпусе из 160 млн. предложений оказывалось меньше 10, то при построении данной дистрибутивной модели порог вхождения в её словарь был опущен до 7.

Еще раз отметим, что данная векторная модель включает в себя и знаки препинания как элементы её словаря. В частности, косинусное расстояние между точкой («.») и восклицательным знаком («!») составило 0.6, что позволяет различать схожие контексты прямого и переносного использования идиоматических выражений, различающиеся знаками препинания (см. пример 3 из раздела 1).

2.3. Параметры и результаты эксперимента

В качестве средства классификации мы использовали в наших экспериментах нейронные сети двух видов:

- (1) простые рекуррентные (Simple RNN);
- (2) двунаправленные рекуррентные (Bidirectional).

В качестве основы программной реализации использовалась библиотека *Keras*.

Сеть состояла из

- (1) входного слоя, принимающего последовательности длины 5 (размер локального контекста), где каждый элемент последовательности есть вектор, представляющий слово или целевое выражение;
- (2) рекуррентного слоя (SimpleRNN или Bidirectional) с функцией активации ReLu (Rectified Linear Unit);
- (3) обычного, полносвязного (*dense*) слоя с функцией активации *softmax*.

Количество нейронов в рекуррентном слое составило 20, а в выходном, производящем непосредственную классификацию, имелось два нейрона.

При проведении экспериментов варьировалось два параметра:

- (1) начальные значения весов сети, что достигалось повторными запусками процесса обучения и тестирования со случайным начальным назначением весов;
- (2) количество итераций обучения изменялось от 20 до 50.

Поскольку данная сеть не относится к классу «глубоких сетей» (*deep networks*), то в экспериментах с ней не применялись средства регуляризации (типа L1, L2 или *dropout*), а также алгоритмы оптимизации скорости обучения (типа *Adam*-оптимизатора).

Эксперименты проводились путем построения отдельного классификатора для каждой из идиом, и в качестве значения точности распознавания бралась максимальная точность, достигнутая при повторных запусках на обучение и тестирование. Множество тестовых примеров получено их случайной выборкой из общего множества примеров использования идиоматического выражения в прямом или переносном смысле. Полученные результаты по точности классификации представлены в следующей таблице.

ТАБЛИЦА 1. *Точность распознавания идиоматических выражений*

	Леммы		Словоформы	
	simple RNN	Bidirectional	simple RNN	Bidirectional
<i>правая рука</i>	94.8	94.8	94.8	95.6
<i>темный лес</i>	93.0	93.0	94.4	95.8
<i>серая мышь</i>	86.9	86.4	92.1	86.4
<i>вчерашний день</i>	91.7	90.3	93.2	92.2
<i>тяжелый груз</i>	84.4	84.0	86.8	86.8
<i>длинный язык</i>	85.1	82.4	86.4	86.9
<i>грязные руки</i>	80.7	82.1	83.5	87.3
<i>красная линия</i>	93.8	93.4	94.3	95.6
<i>закрытая дверь</i>	92.8	93.4	96.5	96.1
<i>последнее слово</i>	87.6	84.5	91.2	90.3

Как видно из приведенной таблицы, использование словоформ, по сравнению со стандартными формами (леммами) слов и выражений, дает в каждом случае более высокие показатели точности распознавания. С другой стороны, применение нейронных сетей того или иного вида — простой рекуррентной или двунаправленной рекуррентной — не дает видимого преимущества для какого-либо вида нейросети в общем случае. Более сложные нейронные сети, например с LSTM-ячейками, не применялись, поскольку их преимущества проявляются при обработке «длинных» последовательностей, например, полных предложений, тогда как в нашем исследовании мы ограничились последовательностями длины 5.

3. Специфика отдельных идиом

Тем не менее, каждая из десяти идиом, с которыми мы проводили эксперименты, обладает своими собственными специфическими особенностями, которые требуют дальнейшего более глубокого изучения. Эти особенности необходимо учитывать при построении систем классификации, подобных нашей, чтобы повысить точность распознавания по сравнению с достигнутой на данный момент. Таким образом, мы предполагаем возможность построения более мощных систем классификации использования идиоматических выражений в виде гибридных систем, объединяющих нейронные сети и некоторые дополняющие

их средства. Именно те примеры использования идиоматических выражений, которые неправильно классифицируются современными нейронными сетями, представляют наибольший интерес для изучения их свойств и особенностей.

Ниже мы кратко перечисляем специфические особенности идиоматических выражений, представленных в разделе 2.1, которые либо уже учтены в нашей реализации классификатора, либо которые необходимо учитывать в более точных системах классификации.

- (1) *правая рука* — почти в 100% случаев при переносном использовании этого выражения оно входит в предложение в виде фрагмента *правая рука* + <имя собственное>; «факторизация» собственных имен до имен классов (в данном случае, класса PERSON) позволяет обрабатывать без потерь редкие собственные имена, которые обычно отсутствуют в словаре дистрибутивной модели, и, тем самым, поднимать до более высокого, семантического уровня обучающие/тестовые примеры, что, в свою очередь, сказывается на повышении точности распознавания.
- (2) *темный лес* — некоторые словоформы этого выражения оказываются очень редко используемыми, а потому, как следует из наших экспериментов, для построения качественной векторной модели слов и выражений требуются корпуса текстов с 1 млрд. и более предложений.
- (3) *серая мышь* — аналогично предыдущей идиоме, в использованном нами корпусе PaRuS оказалось слишком мало предложений с прямым использованием этого словосочетания; таким образом, и в этом случае имеется необходимость в корпусах большего объема.
- (4) *вчерашний день* — использование данного выражения в переносном смысле иногда подчеркивается применением кавычек различного типа (одинарных, двойных, «ёлочек» и т.п.); кроме того, оно может входить в состав более крупных идиом — *искать вчерашний день*, или метафорических определений — *человек вчерашнего дня*.
- (5) *тяжелый груз* — используется в более длинных идиоматических выражениях — *лечь тяжелым грузом (на плечи/сердце)*.
- (6) *длинный язык* — используется в метафорических выражениях *длинные языки пламени/огня, длинные языки подтаявшего снега*, а также входит в состав более крупных идиом и поговорок — *слишком/излишне длинный язык, длинный язык до добра не доведет*.

- (7) *грязные руки* — входит в состав устойчивых выражений типа *Уберите от меня ваши грязные руки!*; часто употребляется с восклицательным знаком на конце выражения; входит в состав метафорических выражений — *За прошедшие годы патриотизм захватили грязными руками.*
- (8) *красная линия* — в 70% случаев использования выражения в переносном смысле, оно заключается в кавычки разного типа; это выражение как идиома близка к выражению *красная черта*; входит в состав часто употребляемого оборота *пересекать красную линию.*
- (9) *закрытая дверь* — чаще всего в переносном смысле используется в выражении *при закрытых дверях*; в сложных случаях, для определения вида использования этого выражения требуется семантическое представление всего предложения, в которое оно входит (см. об этом подробнее в разделе 5).
- (10) *последнее слово* — чаще всего, в переносном смысле используется в выражении *оставлять последнее слово за собой*; также употребляется в составе оборотов *обзывать/ругать последними словами, оборудовать по последнему слову (техники), последнее слово подсудимого.*

4. Смежные работы

Фактически первой работой, в которой было применено векторное представление слов и выражений для задачи распознавания прямого/переносного смысла идиоматических выражений, была работа [4]. Данная задача из списка нескольких задач, решение которых представлено в этой статье, решалась только для одного немецкого идиоматического выражения *ins Wasser fallen* — *падать в воду/терпеть неудачу* — с 37 идиоматическими и 27 прямыми использованиями этого выражения. Сам классификатор строился на основе метода «латентного семантического анализа» (Latent Semantic Analysis).

В соответствии с этим методом, для получения векторного представления слов и целевого идиоматического выражения в том или ином виде его использования, на основе корпуса текстов из немецких газет, строилась матрица совместных вхождений (co-occurences) слов. Строки этой матрицы соответствовали 20000 наиболее часто употребляемым в корпусе словам, а столбцы — 1000 так называемым «задающим содержание» (content-bearing) словам. Под «задающими

содержание» словами понимаются также наиболее часто используемые слова (за исключением стоп-слов), но взятые с некоторым смещением относительно начала списка таких слов, упорядоченного по частоте встречаемости в корпусе.

Далее применялась процедура сингулярного разложения (*Singular Value Decomposition*) для сокращения размерности (а именно, количества столбцов) этой матрицы, которая, в конечном итоге, давала векторное представление слов и выражений длины 100. Различие контекстов, в которых встречались примеры прямого и переносного использования идиомы *ins Wasser fallen*, приводило к тому, что соответствующие векторные представления этой идиомы сильно различались — значение косинусной меры между ними составило 0.02. Для тестового примера использования этой идиомы векторное представление строилось аналогичным образом. Далее оценивалась близость его до полученных ранее векторов прямого и переносного использования идиомы.

Из-за очень маленького размера корпуса (64 предложения), для оценки точности распознавания применялась 10-кратная кросс-валидация на этом корпусе. Средняя точность распознавания целевого выражения *ins Wasser fallen* составила 72%.

В работе [2] предлагается семантический подход к распознаванию прямого/переносного использования идиоматических выражений. Основой подхода является алгоритм, который определяет насколько хорошо прямая (буквальная) интерпретация заданного выражения соответствует «связной структуре текста/дискурса» (*cohesive structure of the discourse*). Если найдены сильные связи (*strong links*) с такой структурой, то выражение классифицируется как используемое в прямом смысле, иначе — в переносном.

Базовым понятием, на которое опирается этот алгоритм, является понятие «лексической цепочки». Под лексической цепочкой авторы понимают последовательность слов, семантически связанных между собою. Поэтому основную идею распознавания того или иного использования идиоматического выражения можно переформулировать так — если отдельные слова целевого выражения не участвуют в лексических цепочках окружающего их текста, то, вероятно, что это выражение используется как идиома. (Пример — если рядом с выражением *правая рука* отсутствуют связанные с ним понятия *палец*,

ладонь, кисть, запястье и т.п., то это выражение будет трактоваться как идиома). Тем самым, в искомым лексических цепочках, используемых для классификации, отображается контекст целевого выражения. Однако, сами авторы этого исследования отмечают, что подобные лексические цепочки существуют только для, приблизительно, 90% исследованных ими идиоматических выражений в прямом их использовании. В качестве меры семантической близости слов использовалось «нормализованное Google-расстояние» (Normalized Google Distance), которое применяется поисковым механизмом Google для ранжирования Web страниц.

В данной работе были исследованы 17 идиом с корпусом из 4000 предложений их использования. Точность распознавания с применением указанного метода составила от 78% до 80.5%. Интересно, что авторы данной работы, кроме подхода на основе лексических цепочек, применили в качестве отдельного, метод обучения с учителем, сходный с методом из работы [4]. Отличие состояло в том, что вместо использования векторного представления слов и выражений, авторы работы использовали прямое пересечение слов тестового предложения с примерами предложений прямого/переносного использования целевого выражения из обучающего множества, что дало точность распознавания 95.69%.

Аналогично работе [2], в работе [9] делается попытка использовать понятие «семантической совместимости» (*semantic compatibility*) идиоматического выражения и его контекста (в работе [2] это понятие названо «семантической связностью» (*semantic relatedness*)). Указанная семантическая совместимость основана на простом наблюдении: прямое значение идиоматического выражения в предложении, где оно употребляется в переносном смысле, до некоторой степени семантически противоречит контексту. Для реализации этой идеи авторы работы [9] предлагают использовать модификацию стандартной дистрибутивной Continuous-Bag-of-Words (CBOW)-модели [10] для получения специальной метрики семантической совместимости прямого значения идиомы и её контекста.

Эта модификация является попыткой преодолеть базовые ограничения CBOW-модели: 1) в ней не учитывается порядок слов в контексте, 2) все слова контекста считаются эквивалентными, хотя некоторые из них могут нести ключевую информацию о семантической совместимости целевого слова и контекста, 3) одно и то же слово

в разных контекстах может иметь разные значения. На самом деле, все эти ограничения сняты в более современных BERT-моделях, которые, в частности, используются для решения задачи распознавания вида использования идиоматических выражений (см. ниже обзор работы [11]).

Технически модификация стандартной CBOW-модели заключалась в использовании 1) нейронов типа LSTM вместо обычных нейронов, 2) слоев глобального и локального внимания (*attention layers*) и 3) сети многослойных персептронов для реализации слоя оценки семантической совместимости (*semantic compatibility evaluation layer*). Обучение такой модели проводилось на корпусе, в качестве которого была взята версия Википедии на английском языке, ограниченная объемом в 100 млн. токенов. Поскольку такой подход к обучению относится к классу несупервизорного обучения (т.е., обучения без учителя), то соответствующая система распознавания вида использования идиоматических выражений является универсальной — нет необходимости обучать систему распознаванию каждой отдельной идиомы.

Авторы статьи сообщают, что полученная система достигает точности распознавания в 74% и значения F-показателя в 76%. Стоит заметить, что эти результаты значительно уступают показателям систем, обученных для распознавания отдельных идиом (см. нашу работу и работу [11]).

В работе [12] также проводится мысль, что различие в прямом и переносном толковании идиоматических выражений можно выявить с использованием их (локальных) контекстов. В свою очередь, для слов контекста и для целевого идиоматического выражения используются векторные представления. Конкретно, в данной работе они были получены на основе известного корпуса text8⁶ со словарем из 71 290 слов. Длина использованных векторов составила 200. Авторы исследовали два метода различения идиоматического и прямого использования выражений:

- (1) Суть первого метода состояла в применении вычисления скалярного (внутреннего) произведения векторов, представляющих контекстные слова, с вектором, представляющим целевое выражение; здесь идея заключалась в том, что при прямом использовании

⁶<http://matmahoney.net/dc/text8.zip>

идиоматического выражения получаемые значения таких произведений будут больше, чем в случае переносного использования, в котором перемножаемые вектора более «перпендикулярны» друг другу;

- (2) Суть второго метода состояла в вычислении ковариационных (**scatter/ covariance**) матриц для векторов, соответствующих словам локального контекста целевого выражения (используемого в том или ином смысле). Получаемые матрицы, как утверждают авторы, представляют распределение слов, составляющих локальный контекст (**local context distribution**). Имея такие матрицы для контекстов прямого и переносного использования идиомы, полученные на основе обучающего множества предложений, их можно сравнивать с аналогичной матрицей, полученной на основе тестового предложения. Расстояние между матрицами измерялось посредством фробениусовой нормы и спектральной нормы, где первая из них дала более лучшие результаты.

В описанном эксперименте было проверено 17 идиоматических выражений со структурой «глагол+существительное», типа *blow whistle*, *lose head* и др. Общее количество использованных обучающих и тестовых примеров составило от 40 до 420 для различных идиом. Точность распознавания составила, в зависимости от конкретной идиомы, от 58% до 87%.

Работа [13] является единственной известной нам работой, в которой задача распознавания прямого/переносного использования идиоматических выражений решается для русского языка. Сам подход применяемый для этого аналогичен подходу, представленному в уже упомянутой выше статье [12]. Подготовленный в рамках этой работы аннотированный корпус русских идиом с их прямым/переносным использованием описан в работе [5].

В работе [11] многословные выражения, которые могут использоваться как в прямом, так и в переносном смысле, названы «потенциально идиоматическими выражениями». Для классификации способов их использования в этой работе предлагается применить «контекстуальные эмбединги» (**contextual embeddings**) [14, 15]. Именно это средство представляет в данном подходе контексты потенциально идиоматических выражений, что позволяет классифицировать последние по способу их использования. Каждое целевое многословное выражение представляется тогда средним значением контекстуальных

эмбедингов отдельных слов этого выражения. В качестве классификатора используется простой однослойный перцептрон, обученный на контекстуальных эмбедингах потенциально идиоматического выражения — эмбедингах для прямого использования выражения и эмбедингах для использования выражения в переносном смысле.

Данный подход испытывался на трех корпусах — на корпусе VNC (verb-noun combinations) и корпусе соревнования SemEval 5b (2013 г.) для английского языка, и корпусе Андреи Хорбах (Andrea Horbach) для немецкого языка. Количество протестированных потенциально идиоматических выражений в этих корпусах составило от 6 до 12, а количество предложений с такими выражениями — от 700 до 5000. Точность распознавания выражений составила от 91% до 94%.

В работе [16] описаны результаты обширного исследования способов распознавания прямого/переносного использования идиоматических выражений с применением различных подходов и корпусов текстов. К корпусу Corpus of Contemporary American English (COCA), размеченному вручную выделением использованных в переносном смысле слов в 14 целевых выражениях, применялась стандартная процедура word2vec для построения векторных представлений слов. Полученные представления соединялись в векторное представление всего выражения тремя способами:

- (1) покомпонентное сложение векторов;
- (2) покомпонентное умножение векторов;
- (3) алгоритм предсказания.

Каждое целевое выражение обретало при этом эталонные векторные представления для прямого и для переносного использования. «Композиционные» вектора для выражений и предложений строились из векторных представлений только существительных, глаголов и прилагательных, входящих в них.

Векторные представления тестовых предложений с идиомой формировались на основе контекста, получающегося исключением целевого выражения из тестового предложения. Представление контекста получалось из векторных представлений слов этого контекста разными способами — от суммирования векторных представлений отдельных слов до выделения в контексте характеризующих слов («ключей»).

Распознавание использования идиомы в тестовом предложении осуществлялось по наибольшему значению косинусной меры между

тестовым представлением и каждым из эталонных. Средняя точность классификации «прямое/переносное использование» составила 82.2%.

5. Выводы и направления дальнейшей работы

В данной работе мы попытались показать, что можно построить достаточно эффективную систему классификации использования идиоматических выражений в прямом/переносном смысле, используя

- (1) локальные контексты;
- (2) исходные словоформы идиом;
- (3) порядок слов в предложении;
- (4) знаки препинания.

Основными компонентами такой системы классификации были простые рекуррентные нейронные сети, а также их двунаправленные варианты, совместно с дистрибутивной системой представления слов и целевых идиоматических выражений. В зависимости от конкретного вида выражения, точность классификации составила от 86% до 96%. Дальнейшее повышение точности распознавания может быть достигнуто, по нашему мнению, применением, в частности, нелинейных (синтаксических) контекстов.

Пример 6. В частности, **красной линией** в комментариях различных экспертов проходит жесткая позиция Саудовской Аравии.

В этом предложении, слово «проходит», которое во многих случаях характеризует переносное использование идиоматического выражения *красная линия*, далеко (в лексическом смысле) отстоит от самой идиомы. Если же брать синтаксическую окрестность (контекст) этой идиомы, то в ней сама идиома и предикат «проходит» будут соседними элементами. В сложных случаях, для определения способа использования идиоматического выражения может помочь (автоматически извлекаемая) некоторого рода семантическая структура предложения, включающего это выражение.

Пример 7. Рэйчел смотрела на **закрытые двери**. Переговоры происходили за **закрытыми дверями**.

В первом предложении выражение *закрытые двери* используется в прямом смысле. Помочь это определить может семантическая структура этого предложения вида:

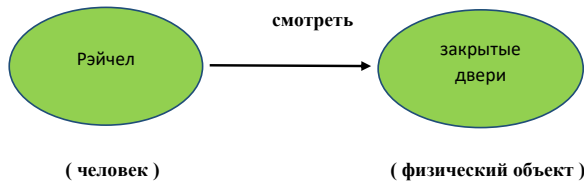


РИСУНОК 2. Семантическое представление предложения

При наличии средств распознавания событий/ситуаций, в которых (закрытая) дверь участвует как физический объект во взаимодействии с другими объектами, упоминаемыми в предложении, мы сможем точнее определять прямое использование выражения *закрытая дверь*. Здесь, в частности, может оказаться полезной теория семантических ролей Ч. Филмора [17] (см. также [18]).

Список литературы

- [1] A. Fazly, P. Cook, S. Stevenson. “Unsupervised type and token identification of idiomatic expressions”, *Computational Linguistics*, **35**:1 (2009), pp. 61–103.  [↑]₄
- [2] L. Li, C. Sporleder. “Classifier combination for contextual idiom detection without labelled data”, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Singapore, 6–7 August 2009), ACL, 2009, pp. 315–323.  [↑]_{4,16,17}
- [3] Ю. Г. Бадрызлова. *Автоматические методы распознавания метафоры в текстах на русском языке*, Дис. ... канд. фил. наук, НИУ «Высшая школа экономики», М., 2019 (in English), 206 с.  [↑]₄
- [4] G. Katz, E. Giesbrecht. “Automatic identification of non-compositional multiword expressions using Latent Semantic Analysis”, *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties* (Sydney, July 2006), ACL, 2006, pp. 12–19.  [↑]_{5,15,17}
- [5] K. Aharodnik, A. Feldman, J. Peng. “Designing a Russian idiom-annotated corpus”, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018* (Miyazaki, Japan, 7–12 May 2018), ELRA, 2018, pp. 2533–2538.  [↑]_{5,19}
- [6] D. Puzyrev, A. Shelmanov, A. Panchenko, E. Artemova. “Noun compositionality detection using distributional semantics for the Russian language”, *Analysis of Images, Social Networks and Texts*, AIST 2019, Lecture Notes in

- Computer Science, vol. **11832**, eds. van der Aalst W. et al., Springer, Cham, 2019, ISBN 978-3-030-37333-7, pp. 218–229.  [↑]₆
- [7] Anh Le The, M. Burtsev. “A deep neural network model for the task of named entity recognition”, *International Journal of Machine Learning and Computing*, **9**:1 (2019), pp. 8–13.  [↑]₈
- [8] Н. А. Власова, И. В. Трофимов, Ю. П. Сердюк, Е. А. Сулейманова, И. Н. Воздвиженский. «PaRuS — синтаксически аннотированный корпус русского языка», *Программные системы: теория и приложения*, **10**:4(43) (2019), с. 181–199.    [↑]₁₀
- [9] C. Liu, R. Hwa. “A generalized idiom usage recognition model based on semantic compatibility”, *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**:01 (2019), pp. 6738–6745.  [↑]₁₇
- [10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean. “Distributed representations of words and phrases and their compositionality”, *Proceedings of the 26th International Conference on Neural Information Processing Systems*. V. 2, NIPS’13 (Lake Tahoe, Nevada, USA, December 5–10, 2013), Curran Associates Inc., 2013, pp. 3111–3119.  [↑]₁₇
- [11] M. Kurfali, R. Ostling. “Disambiguation of potentially idiomatic expressions with contextual embeddings”, *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons* (Barcelona, Spain (Online), December 13, 2020), ACL, 2020, pp. 85–94.  [↑]_{18,19}
- [12] J. Peng, A. Feldman, H. Jazmati. “Classifying idiomatic and literal expressions using vector space representations”, *Proceedings of the International Conference Recent Advances in Natural Language Processing* (Hissar, Bulgaria, Sep 7–9, 2015), INCOMA Ltd., 2015, pp. 507–511.  [↑]_{18,19}
- [13] J. Peng, K. Aharodnik, A. Feldman. “A distributional semantics model for idiom detection — the case of English and Russian”, *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*. V. 2 (Funchal, Madeira, Portugal, Jan 16–18, 2018), SciTePress, 2018, ISBN 9789897582752, pp. 675–682.  [↑]₁₉
- [14] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer. “Deep contextualized word representations”, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. V. 1 (New Orleans, Louisiana, USA, June 1–6, 2018), ACL, 2018, pp. 2227–2237.   [↑]₁₉
- [15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. “BERT: Pre-training of deep bidirectional transformers for language understanding”, *Proceedings of NAACL-HLT 2019*. V. 1 (Minneapolis, Minnesota, USA, June 2–June 7, 2019), ACL, 2019, pp. 4171–4186.  [↑]₁₉
- [16] F. Sa-Pereira. *Distributional representations of idioms*, Masters Thesis, McGill University, 2016, 109 pp.  [↑]₂₀

- [17] Ch. Fillmore. “The case for case”, *Universals in Linguistic Theory*, eds. E. Bach, R. Harms, Holt, Rinehart, and Winston, 1968, pp. 21–119.  [↑]₂₂
- [18] A. Shelmanov, I. Smirnov. “Methods for semantic role labeling of Russian texts”, *Computational Linguistics and Intellectual Technologies*, **13**:20, Proceedings of International Conference Dialogue (2014), pp. 607–620.  [↑]₂₂

Поступила в редакцию 19.07.2021

Переработана 17.09.2021

Опубликована 27.09.2021

Рекомендовал к публикации

д.ф.м.-н. Н. В. Лукашевич

Пример ссылки на эту публикацию:

Ю. П. Сердюк, Н. А. Власова. «Распознавание идиоматического использования выражений с помощью нейронных сетей». *Программные системы: теория и приложения*, 2021, **12**:3(50), с. 3–27.

 10.25209/2079-3316-2021-12-3-3-27

 http://psta.psiras.ru/read/psta2021_3_3-27.pdf

Об авторах:



Юрий Петрович Сердюк

старший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна, научные интересы: параллельное программирование, формальные исчисления процессов, системы типов.



0000-0003-2916-2102

e-mail: Yuri@serdyuk.botik.ru



Наталья Александровна Власова

младший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, одна из разработчиков технологии построения систем извлечения информации.



0000-0002-7843-6870

e-mail: nathalie.vlassova@gmail.com

CSCSTI 16.31.21
UDC 004.89:81'322.2

Yuri Serdyuk, Natalia Vlasova. *Idiomatic expression usage recognition by neural networks.*

ABSTRACT. Many of the idiomatic expressions can be used both in literal and non-literal ways. The recognition of such cases is an important problem in many natural language processing applications, namely, in machine translation. We propose automatic idiom usage recognition method based on the analysis of local contexts of such expressions. We apply recurrent neural networks to solve this problem. Two types of neural networks are investigated — simple and bidirectional recurrent networks. We compare two forms of representation of context words — the canonical form (by lemmas) and by source word forms. We describe construction and parameters of the distributive model which stores the vector representations of single words and target idiomatic expressions. Due to the great diversity of approaches to solving the idiom usage recognition problem, we provide an extended survey of basic efforts in this domain.

2020 *Mathematics Subject Classification:* 68T50; 68T05, 91F20

References

- [1] A. Fazly, P. Cook, S. Stevenson. “Unsupervised type and token identification of idiomatic expressions”, *Computational Linguistics*, **35**:1 (2009), pp. 61–103.  ⁴
- [2] L. Li, C. Sporleder. “Classifier combination for contextual idiom detection without labelled data”, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Singapore, 6–7 August 2009), ACL, 2009, pp. 315–323.  ^{4, 16, 17}
- [3] Yu. G. Badryzlova. *Automated metaphor identification in Russian texts*, Dis. ... kand. fil. nauk, NIU “Vysshaya shkola ekonomiki”, M., 2019 (in English), 206 pp.  ⁴
- [4] G. Katz, E. Giesbrecht. “Automatic identification of non-compositional multiword expressions using Latent Semantic Analysis”, *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties* (Sydney, July 2006), ACL, 2006, pp. 12–19.  ^{5, 15, 17}
- [5] K. Aharodnik, A. Feldman, J. Peng. “Designing a Russian idiom-annotated corpus”, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018 (Miyazaki, Japan, 7–12 May 2018), ELRA, 2018, pp. 2533–2538.  ^{5, 19}
- [6] D. Puzyrev, A. Shelmanov, A. Panchenko, E. Artemova. “Noun compositionality detection using distributional semantics for the Russian language”, *Analysis of Images, Social Networks and Texts*, AIST 2019, Lecture Notes in Computer Science, vol. **11832**, eds. van der Aalst W. et al., Springer, Cham, 2019, ISBN 978-3-030-37333-7, pp. 218–229.  ⁶

- [7] Anh Le The, M. Burtsev. “A deep neural network model for the task of named entity recognition”, *International Journal of Machine Learning and Computing*, **9**:1 (2019), pp. 8–13.  ⁸
- [8] N. A. Vlasova, I. V. Trofimov, Yu. P. Serdyuk, Ye. A. Suleymanova, I. N. Vozdvizhenskiy. “PaRuS — syntax annotated Russian corpus”, *Program Systems: Theory and Applications*, **10**:4(43) (2019), pp. 181–199 (In Russian).   ¹⁰
- [9] C. Liu, R. Hwa. “A generalized idiom usage recognition model based on semantic compatibility”, *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**:01 (2019), pp. 6738–6745.   ¹⁷
- [10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean. “Distributed representations of words and phrases and their compositionality”, *Proceedings of the 26th International Conference on Neural Information Processing Systems*. V. 2, NIPS’13 (Lake Tahoe, Nevada, USA, December 5–10, 2013), Curran Associates Inc., 2013, pp. 3111–3119.  ¹⁷
- [11] M. Kurfali, R. Ostling. “Disambiguation of potentially idiomatic expressions with contextual embeddings”, *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons* (Barcelona, Spain (Online), December 13, 2020), ACL, 2020, pp. 85–94.  ^{18,19}
- [12] J. Peng, A. Feldman, H. Jazmati. “Classifying idiomatic and literal expressions using vector space representations”, *Proceedings of the International Conference Recent Advances in Natural Language Processing* (Hissar, Bulgaria, Sep 7–9, 2015), INCOMA Ltd., 2015, pp. 507–511.  ^{18,19}
- [13] J. Peng, K. Aharodnik, A. Feldman. “A distributional semantics model for idiom detection — the case of English and Russian”, *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*. V. 2 (Funchal, Madeira, Portugal, Jan 16–18, 2018), SciTePress, 2018, ISBN 9789897582752, pp. 675–682.  ¹⁹
- [14] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer. “Deep contextualized word representations”, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. V. 1 (New Orleans, Louisiana, USA, June 1–6, 2018), ACL, 2018, pp. 2227–2237.   ¹⁹
- [15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. “BERT: Pre-training of deep bidirectional transformers for language understanding”, *Proceedings of NAAACL-HLT 2019*. V. 1 (Minneapolis, Minnesota, USA, June 2–June 7, 2019), ACL, 2019, pp. 4171–4186.  ¹⁹
- [16] F. Sa-Pereira. *Distributional representations of idioms*, Masters Thesis, McGill University, 2016, 109 pp.  ²⁰
- [17] Ch. Fillmore. “The case for case”, *Universals in Linguistic Theory*, eds. E. Bach, R. Harms, Holt, Rinehart, and Winston, 1968, pp. 21–119.  ²²
- [18] A. Shelmanov, I. Smirnov. “Methods for semantic role labeling of Russian texts”, *Computational Linguistics and Intellectual Technologies*, **13**:20, Proceedings of International Conference Dialogue (2014), pp. 607–620.  ²²

Sample citation of this publication:

Yuri Serdyuk, Natalia Vlasova. “Idiomatic expression usage recognition by neural networks”. *Program Systems: Theory and Applications*, 2021, **12**:3(50), pp. 3–27. (In Russian).

 10.25209/2079-3316-2021-12-3-3-27

 http://psta.psiras.ru/read/psta2021_3_3-27.pdf