Федеральное государственное автономное образовательное
учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Образовательная программа
01.03.02 Прикладная математика и информатика

ВЫПУСКНАЯ КВАЛИФАЦИОННАЯ РАБОТА

на тему

Методы обучения с учителем для определения некомпозициональных именных групп

Выполнил студент гр. БПМИ 154, 4 курса
Пузырев Дмитрий Александрович

Научный руководитель:
Артемова Е.Л., к.т.н.

Москва, 2019
2018/2019 уч.г.

Содержание

Аннотация

Композициональность именных групп является индикатором того, насколько точно смысл всей фразы может быть описан через семантические свойства её частей и их грамматическую связь. В случае, если семантическое значение сочетания не совпадает со значениями его компонент, оно называется некомпозициональным. Примерами некомпозициональных фраз могут являться такие идиомы, как "бархатная революция"и "тёмная лошадка". Умение выделять такие фразы принесет ощутимую пользу в ряде прикладных задач, в частности, в машинном переводе. Ранее проблема определения композициональности решалась с помощью применения математических операций с векторами слов, полученных из моделей дистрибутивной семантики. В отличие от предыдущих работ, напрямую не использующих концепт обучения, мы формулируем эту задачу как классификацию и представляем различные известные алгоритмы обучения с учителем, получая значимое улучшение качества предсказаний на известных примерах по сравнению с прошлыми методами. Также представлен датасет именных групп структуры прил.-сущ. и сущ.-сущ. для русского языка, первый подобный для славянских языков. С помощью аналогичных методов классификации мы получаем базовые результаты качества предсказания, сравнимые с англоязычными аналогами.

Abstract

Compositionality of noun compounds indicates to what extent the meaning of a phrase can be derived from the meaning of its parts and their grammatical relations. If semantics of a compound differs from its own components, the phrase is called non-compositional. Some of the examples of non-compositional collocations are "hot dog"and "rat race". This can be applied in various areas, notably, in mashine translation. Detection of compositionality is a task that has been frequently addressed with Distributional Semantic Models (DSMs) and mathematical operations with corresponding word vectors. Unlike traditional approaches that use an "unsupervised setting we treat this task as a classification problem and introduce various supervised learning algorithms to gain substantial improvement across gold standard dataset over state-of-the-art models. We also proceed to introduce Russian language dataset for compound compositionality task with ADJ+NOUN and NOUN+NOUN speech patterns, the first one among Slavic languages to our knowledge, and perform same classification methods to introduce baseline results comparable with english corpora.

1 Введение

Одной из наиболее важных прикладных задач в области обработки естественного языка (*NLP*) является создание представлений структурных единиц языка (слово, словосочетание, предложение и.т.п.), учитывающих и отражающих их смысловое наполнение. Большое количество работ посвящено созданию различных моделей дистрибутивной семантики (напр. *Word2vec* [1], *ELMo* [2]), представляющих из себя векторное представление смысловой компоненты. Наиболее распространенным способом получения представлений синтаксических элементов из их атомарных единиц (к примеру, предложений из слов) до сих пор является усреднение компонент, либо элементарное, либо взвешенное с помощью какой-либо метрики. Несмотря на то, что такой подход часто дает приемлемое представление необходимого элемента, хотелось бы использовать методы, основанные на принципах семантики слов и фраз.

В семантике существует понятие композиции, согласно которой смысл единицы в общем случае может быть каким-либо образом выведен из представлений иерархически более низких элементов, к примеру, смысл фразы "вагон поезда"является ровно суммой смыслов слов "вагон"и "поезд". Однако, для целого ряда фраз такая связь не может быть установлена напрямую. Чаще всего такие фразы являются идиомами — их семантическое значение неделимо. Примерами могут послужить такие фразы, как "бархатная революция"или "тёмная лошадка". Такие представления могут быть прослежены, в частности, в работе лингвиста В.В.Виноградова "Об основных типах фразеологических единиц в русском языке"[3]. Он предлагает разделять фразелогизмы на три категории: *фразелогические сочетания*, *фразелогические единства* и *фразелогические сращения*. *Фразеологические сочетания* состоят из слов со свободным значением и характеризуются семантической разложимостью (*верный друг*, *трескучий мороз*).Во *фразеологических единствах* имеется общее переносное значение, но отчетливо сохраняются признаки семантической разложимости компонентов (*кровь с молоком*, *стрелянный воробей*), В то время как о *семантических сращениях* Виноградов пишет

> *Несомненно, что легче и естественнее всего выделяется тип словосочетаний - абсолютно неделимых, неразложимых, значение которых совершенно независимо от их лексического состава,*

от значений их компонентов и так же условно и произвольно, как значение немотивированного слова-знака.

В данной работе предложены решения задачи определения композициональности. Она состоит в нахождении численной оценки, отражающей, насколько смысл фразы выводим из смыслов её слов-компонент.

Предыдущие работы решали данную задачу для примеров на английском и некоторых других западноевропейских языках. Основным предположением, на котором основывались описанные методы, была гипотеза о большей близости векторных представлений по тем или иным метрикам для композициональных фраз по сравнению с некомпозициональными.

Мы формулируем задачу в несколько другом виде. Будем считать, что композициональность именной группы может быть представлена опеределенным классом композициональности. В простейшем случае определим два класса (0, если фраза некомпозицональна, 1, если композициональна) и поставим целью предсказать на основе векторных представлений именной группы и слов-компонент по отдельности, к какому классу относится фраза. Тогда проблема определена как задача бинарной классификации, где координаты векторных представлений элементов будут являться признаками.

Определение композициональности фразы потенциально имеет важность для ряда задач обработки естественного языка. В области *машинного перевода* такой подход помогает выделять семантически неделимые единицы в тексте и переводить их как единое целое вместо попыток перевести компоненты по отдельности. В задаче определения и разделения смыслов (*word sense disambiguation*) некомпозициональным фразам должен сопоставлен один общий смысл. Также очевидны перспективы применения в *семантическом парсинге*.

К сожалению, в открытом доступе отсутствуют ресурсы по композициональности фраз в русских текстах. Поэтому был составлен датасет русских именных групп на основе русских текстов из Universal Dependencies[1]. Тексты из этого ресурса обладают размеченными синтаксическими деревьями, что позволяет выделить нужные синтаксические структуры. Важным отличием от предыдущих датасетов является наличие контекстов для каждого примера, что позволит предсказывать композициональность фразы на основе словесного окружения в будущем.

[1] `https://universaldependencies.org`

В этой работе описан следующий вклад в поставленную задачу:

1. Представлена формулировка задачи предсказани композциональности как классификации; применены различные методы обучения с учителем для определения классов; результаты тестирования сопоставлены с предыдущими экспериментами без обучения

2. Собран и аннотирован датасет по композициональностидля именных групп русского языка; проведены эксперименты на основе алгоритмов обучения с учителем; проведено сравнение качества предсказания с англоязычными аналогами.

2 Обзор релевантной литературы

2.1 Существующие датасеты на иностранных языках

Первые датасеты, отражающие проблему композициональности словосочетаний, появляются в начале 2000-х годов. В работе Baldwin,2002 [4] из корпуса новостей Wall Street Journal извлечены конструкции типа глагол-предлог и даны бинарные оценки того, является ли извлеченное сочетание фразовым глаголом. В следующей статье (Baldwin, 2003) [5] аналогичный метод парсинга использован для извлечения 1710 именных компаундов. В (McCarthy, 2003) [6] 116 примеров английских фразовых глаголов оцениваются тремя аннотаторами по шкале от 0 до 10. (Venkapathy, 2005) [7] используют 800 коллокаций типа "глагол-объект"для получения аннотаций от 1 до 6 где 1 обозначает чистую некомпозициональность и 6 означает чистую композициональность.

В работе (Reddy, 2011) [8] каждой из 90 английских именных групп сопоставлено значение композиональности на основе усреднения 30 оценок. В этой статье представлены значения композициональности как для всей фразы, так и для её составляющих. Это позволяет использовать различные векторные операции с эмбеддингами сочетания и его частей в контексте сопоставления человеческих представлений и семантической близости.

В (Ramisch, 2016) [9] предыдущий датасет был расширен до 180 примеров и созданы параллельные датасеты на французском и португальском языках. Английские паттерны "существительное-существительное"были сопоставлены конструкциям "существительное-предлог-существительное"и "существительное-

прилагательное"согласно грамматическим особенностям языков. В (Farahmand, 2015) [10] публикуется более объемный датасет из 1042 групп "сущ.-сущ. аннотированный 4 экспертами.

Также заслуживают внимания несколько работ с другими языками. В (Gurrutxaga, 2013) [11] 1,200 коллокаций "сущ.-глаг."на баскском языке разбиваются на три класса: идиомы, коллокации и свободные сочетания. 244 немецких сочетания с оценками композициональности от 1 до 7 представлены в (Roller, 2013) [12].

2.2 Алгоритмы решения

Одной из первых попыток определения композициональности были проведены в работе (Baldwin, 2003) [5]. В ней используется LSA для подсчета близости между фразами и его компонентами. В (Venkapathy, 2005) [7] эта идея расширяется добавлением признаками коллокации (частотность фраз, PMI). Попытки изучить возможность вывести семантику компаундов через его части активно изучалась в более ранних работах (McCarthy et al.,2003 [6]).

Предыдущие исследования содержат разные подходы к определению композициональности сочетаний. В основном, они основываются на подходах без применения машинного обучения: различные метрики близости между фразой и её частями сопоставляются с ответами аннотаторов путем замера корреляции Спирмэна ρ.

Базовые результаты были получены Siva Reddy в статье [8]. Кроме представления активно используемого датасета именных групп, авторы исследовали ответы респондентов, в частности, какой вклад атомарные элементы фразы вносят в итоговую композициональность. Примитивные аддитивные и мультипликативные модели на основе эмбеддингов были использованы для исследования. Было установлено наличие существенной корреляции между мерой буквальности слов и композициональностью всей фразы.

В [13] представлен иной подход к проблеме. На эмбеддингах *Word2Vec* обучается отображение из вектора слов в вектор компаунда.

Далее представлен пример линейной функции отображения f для эмбеддингов w_i и w_j размерности d. Решается задача *минимизации* среднеквадратичной ошибки θ.

$$f(\phi(w_i), \phi(w_j)) = [\phi(w_i), \phi(w_j)]\theta_{2d\times d} \quad (1)$$

$$\min_{\theta} |||[\phi(w_i), \phi(w_j)]\theta_{2d\times d} - \phi(w_i, w_j)|| \quad (2)$$

Далее, авторами делается предположение о том, что примеры с маленькой среднеквадратичной ошибкой хорошо отображаются в смысл всей фразы, а значит, являются композициональными. В противном случае пример будет являться некомпозициональным. Таким образом, полученные ошибки используются фактически как те же метрики близости между векторами. Как отмечает автор, такой метод не может напрямую считаться обучением композициональности, так как модель не видит реальной информации по композициональности фраз. Для получения отображения были применены линейная, полиномиальная регрессия и нейронные сети. Была получена корреляция Спирмэна в 0.4103 на датасете *Farahmand*.

Еще один интересный метод без использования обучения представлен в [14]. С помощью TF-IDF ранжирования составлен ранговые списки компаундов и его компонент. Далее, ряд метрик был использован для подсчета расстояния между списками, в частности, расстояния *Манхэттэн*, *Хэмминга*, *Чебышева*, и корреляция Пирсона. Наилучшие результаты были получены с ипсользованием корреляции Пирсона.

Модель	**Корреляция Спирмэна**
ADD (Reddy et al., 2011)	0.21
MULT (Reddy et al., 2011)	0.09
Sparse Interaction polynomial projection (Yazdani et al., 2015)	0.41
Pearson Correlation between ranked lists (Lioma et al., 2017)	0.62

Табл. 2.2.1. Избранные результаты определения композициональности из предыдущих работ на датасете *Farahmand*

3 Сбор и аннотация русских примеров

3.1 Сбор данных

Именные группы были собраны из русских трибанков Universal Dependencies (UD): SynTagRus, GSD, Taiga. В них представлены тексты разных тематик: новости, художественные произведения, публицистика. Поскольку каждому корпусу сопутствует стандартизированная разметка, нет необходимости про-

водить дополнительный POS-тэггинг и построение синтаксичесеих деревьев. Согласно вручную размеченным синтаксическим деревьям были выделены конструкции двух видов:

1. **AN**: существительное именительного падежа и подчинённое ему прилагательное, напр. *новый год*, *открытое море*.

2. **NN**: существительное именительного падежа и подчиненное еу существительное родительного падежа, напр. *точка кипения*, *царь горы*.

Все трибанки парсятся на предмет наличия таких структур. Далее, отбирается 10000 наиболее частотных примеров каждого вида. Из них создается выборка из 1000 случайных фраз (500+500) для аннотации. Для каждой группы верны следующие утверждения:

- приведена к начальной форме с помощью лемматизатора Mystem [15] (т.е. главное существительное приведено к начальной форме, а подчинённое слово согласовано с ним)

- приведена к нижнему регистру (напр. *белый дом* и *Белый дом* считаются одной группой)

- ударение не учитывается (*бóльшая часть* и *большáя часть* — одна группа)

3.2 Аннотация

Каждая из отобранных 1000 фраз оценена двумя аннотаторами по следующей схеме:

- Если фраза некомпозициональна, то она принадлежит классу **0**

- Если фраза композициональна, то она принадлежит классу **1**

- Если фраза многозначна, и её композициональность зависит от контекста, то она принадлежит классу **2**

Далее, ответы аннотаторов просматриваются модератором разметки. Из изначальных 1000 примеров выбирается 220 и разрешаются разногласия в ответах между экспертами.

Метрика	Значение
Корелляция Пирсона	0.541
α Кронбаха	0.700

Табл. 3.2.1. Метрики согласия аннотаторов на тысяче отобранных примеров

В **Таблице 3.2.1.** представлена степень согласия аннотаторов разметки на 1000 примерах.

Была достигнута приемлемая мера согласия в разметке. Можно выделить два основных вида допущенных разногласий:

- Фраза композициональна по определению, но в метафорическом смысле, т.е. является фразеологическим сочетанием по Виноградову (*открытое море*)
- Фраза соедржит полисемию (*ход дела* — судебного или несудебного)

3.3 Описание датасета

Получившийся датасет состоит из 220 именных групп вида *сущ.-сущ.* и *прил.-сущ.*. В **Таблице 3.3.1.** представлена статистика фраз по конструкции и классам. В **Таблице 3.3.2.** даны примеры компаундов каждого класса.

	Прил.-Сущ.	Сущ.-Сущ.	$\sum$
0	23	10	33
1	71	96	167
2	9	11	20
$\sum$	103	117	220

Табл. 3.3.1. Количество композициональных и некомпозициональных примеров в датасете

0	*горячая точка*, *железный занавес*, *каменный век*, *царь горы*, *новая волна*
1	*авиационная бомба*, *гимн страны*, *горнолыжный курорт*, *дно океана*, *федеральный закон*
2	*новый год*, *крупная сеть*, *огромная масса*, *позиция компании*, *древнейшая профессия*

Табл. 3.3.2. Примеры фраз класса 0 (некомпозициональность), класса 1 (композициональность) и класса 2 (многозначность)

Отличием от предыдущих аналогов является наличие контекстов употребления для каждого элемента. Композициональность внутри контекстов на

данный момент не размечена. Примеры контекстов представлены на **Рис. 3.3.3.**

Под воздействием этого поля ядра	**атомов водорода**	в теле исследуемого , которые представляют собой маленькие магнитики , каждый со своим слабым магнитным полем , ориентируются определенным образом относительно сильного поля магнита .
Прозрачная жидкость , в которой на два	**атома водорода**	приходится один атом кислорода , может быть водой , а может быть и смесью жидких водорода и кислорода (внимание : не смешивать в домашних условиях !) .
Нам удалось сложить кучку из восьми атомов - двух атомов углерода и шести	**атомов водорода**	, изображенную на рисунке .
С чего начинать : сдвинуть два атома углерода или приставить	**атом водорода**	к атому углерода ?
Китайский	**Новый год**	и другие праздники , отмечаемые тайскими китайцами , отличаются в обоих случаях , так как они рассчитываются по китайскому календарю .
Перед самым	**Новым годом**	отключили поселок Никольское .
Речь , конечно же , идет об очередной заморозке до	**нового года**	цен на бензин .
В нашем рейтинге лучших подарков мужчине под	**Новый год**	пневматическая винтовка с ночным прицелом твердо заняла первое место .
Нынешнее заседание Госсовета - первое в	**новом году**	и последнее , на котором Владимир Путин выступит как президент страны .
- Но это же был единственный русский фильм на	**Новый год**	, у него были все шансы на успех " .
А у нас политик	**второго эшелона**	ниже этого эшелона не опустится " , - говорит эксперт .
Несмотря на озабоченность Минобрнауки бесконтрольным размножением экономистов и недоверие солидных работодателей к дипломам вузов	**второго эшелона**	, молодой экономист сегодня вряд ли останется на обочине жизни .
Пока потребители	**второго эшелонов**	дожидаются сезона распродаж или приобретают подержанные вещи , лидеры консюмеризма переходят к следующей фазе потребления .
А вот концепция потоковой структуры неравенства дает объяснение их парадоксальной стратегии : опускаясь по стратификационной лестнице , они опережают по статусу тех , кто находится во	**втором эшелоне**	, то есть в предшествующей фазе потребительской гонки .

Рис.3.3.3. Примеры контекстов для компаундов

4 Описание использованных методов

4.1 Методика для английских компаундов

В случае с английскими текстами, мы использовали датасет *Farahmand*, представленный в [10]. Он состоит из 1042 фраз, извлеченных из Википедии, и ответов четырех экспертов по их композициональности. Для последующих задач была использована сумма этих бинарных предсказаний. Таким образом, для каждого примера получили значения от 0 (чистая некомпозициональность) до 4 (чистая композициональность). Для удобства экспериментов мы переформулировали задачу как предсказание величины композициональности с помощью регрессии вместо классификации на 5 классов.

Для устойчивости предсказаний было сделано 25 случайных выборок из датасета. Каждая выборка делилась на обучающие и тестовые примеры в пропорции 75%-25%. Для каждой фразы, признаками является конкатенация координат векторных представлений слов и компаунда целиком. Для обучения рассматривалось несколько простых алгоритмов:

1. Support Vector Regression
2. Kernel Ridge Regression
3. Stochastic Gradient Descent Regression
4. Kernel Ridge Regression
5. K Nearest Neighbours Regression
6. Partial Least Squares Regression

В эксперименте 1 мы изучили, как будут работать предсказания на основе эмбеддингов моделей *Word2Vec*. На основе английской Википедии были обучены эмбеддинги размерности 50. В текстах были выделены компаунды. Таким образом, получили задачу обучения с 150 признаками.

Однако хотелось бы иметь еще какую-то информацию, кроме дистрибутивной. Для этого подойдут эмбеддинги Poincaré, описанные в [16]. Такая модель позволяет обучать *иерархические* репрезентации слов за счёт их отображения в гиперболическое пространство. Такая гиперболическая геометрия позволяет учитывать одновременно *семантическую близость* и *семантическую иерархию*. Под *семантической иерархией* подразумевается, в частности, информация о *гипонимах* и *гиперонимах* слова.

Списки гиперонимов и гпонимов были извлечены из корпуса [17], используя лексико-синтаксические паттерны, описанные в работе [18]:

- such NP as NP, NP[,] and/or NP;
- NP such as NP, NP[,] and/or NP;
- NP, NP [,] or other NP;
- NP, NP [,] and other NP;
- NP, including NP, NP [,] and/or NP;
- NP, especially NP, NP [,] and/or NP;

Далее списки пар "гипоним-гипероним"сортируются по частотности. Обучаются эмбеддинги Poincaré размерности 50.

Для получения оценки композициональности мы берем взвешенную сумму предсказаний двух моделей, основанных на эмбеддингах *Word2Vec* и *Poincaré*

$$Score_S(w_1w_2) = (1-\alpha) * Score_{DS}(w_1w_2) + \alpha * Score_{PS}(w_1w_2) \quad (3)$$

Здесь *DS* представляет модель, обученную на дистрибутивной информации (*Word2Vec*), *PS* — модель, обученную на иерархической информации (*Poincaré*), α — коэффициент пропорции.

Из модели *Poincaré* было получено 780 из 1042 эмбеддингов для фраз. В связи с этим мы предоставляем результаты экспериментов для двух ситуаций: для 780 примеров с восстановленными эмбеддингами (далее **FD-780**) и для всех примеров (далее **FD-1042**). В последнем случае отсутствующий вектор заменяется вектором нулей.

4.2 Методика для русских компаундов

Похожие эксперименты были проведены и для русских фраз. Для чистоты эксперимента, необходимо установить базовые результаты, основанные на применении метрик без обучения. Для этого были использованы инструкции, описанные в [9]. В ней измеряется косинусная близость между компаундом целиком и суммой его частей.

Рассмотрим слова w_1, w_2 и функцию $v(\cdot)$, обозначающую векторное представление слова или группы. Тогда Близость будет измерятся как

$$cos(v(w_1w_2), v(w_1 + w_2)) \quad (4)$$

Где $v(w_1 + w_2)$ является нормализованной суммой векторных представлений:

$$v(w_1 + w_2) = \frac{v(w_1)}{\|v(w_1)\|} + \frac{v(w_2)}{\|v(w_2)\|}. \quad (5)$$

Кроме косинусного расстояния, описанного в этой статье, также были использованы расстояние Чебышева (L_∞-норма), расстояние Манхэттэн (L_1-норма) и Евклидово расстояние (L_2-норма). При использовании этих метрик, вместо нормы применялось усреднение:

$$v(w_1 + w_2) = \frac{1}{2}(w_1 + w_2). \quad (6)$$

Для оценки качества модели подобно релевантным английским статьям использовалась корреляция Спирмэна между значением метрики и оценками аннотаторов в нашем датасете.

Решение задачи с помощью обучения приводится в виде классификации, как было описано во введении. Для этого рассматриваются только примеры с классами 0 и 1 (т.е. те, для которых нет необходимости разрешать многозначность), что оставляет нам 200 фраз. Были использованы следующие алгоритмы классификации:

1. Классификация методом опорных векторов (LSVC) [19], C=1
2. Перцептрон с тремя слоями (MLP) [20] со слоями в 200/20/20 нейронов
3. Дерево решений (DT) [21] глубиной в 10
4. Метод наивного Байеса (NB) [22]

Была также использована сверточная нейронная сеть с одним одномерным сверточным слоем. Для нее данные разделяются на 3 канала: эмбеддинг слова 1, эмбеддинг слова 2, эмбеддинг компаунда. Также в ней содержатся 2 feed-forward слоя на 120 и 84 нейрона. На выхде применяется softmax и выдаются вероятности класса. В качестве функции потерь используется recall класса 0 — таким образом, сеть учится искать некомпозициональные примеры. Схематичное изображение архитектуры представлено ниже.

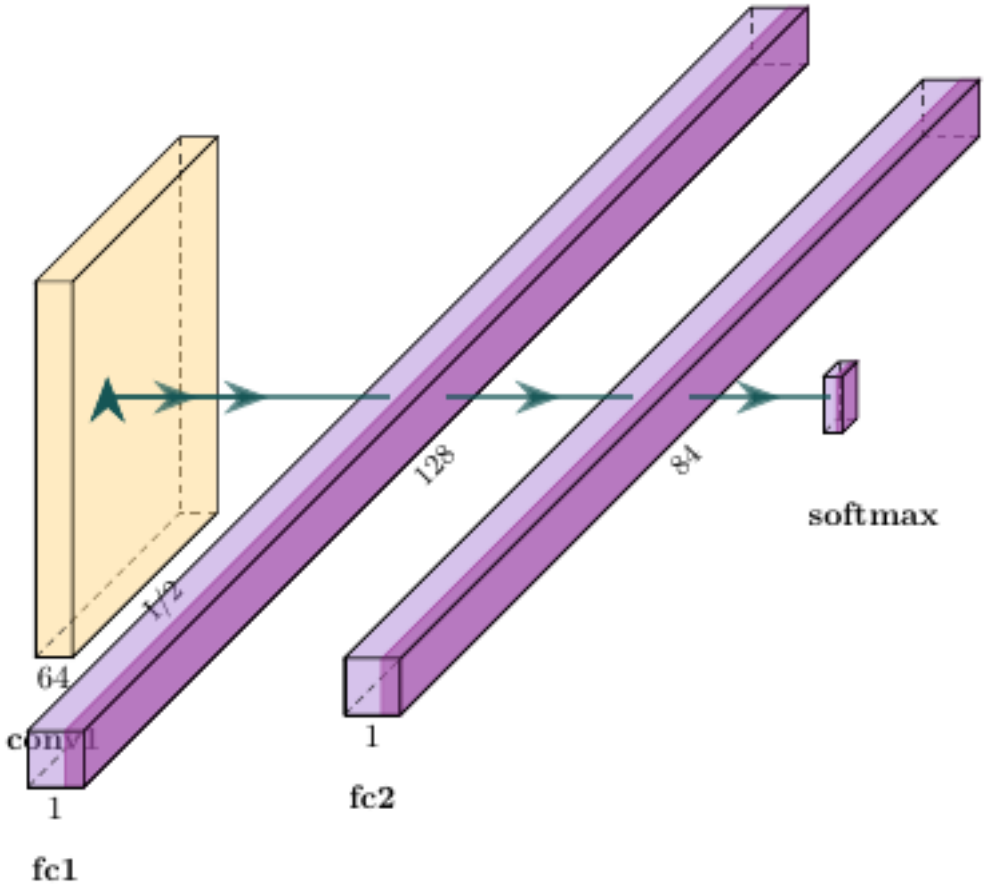

Рис.4.2.1. Архитектура сети.

Как и в случае с английскими компаундами, в качестве признаков используется конкатенация векторов слов и вектора компаунда.

Для обучения дистрибутивной информации использовались эмбеддинги FastText [?]. Также были использованы эмбеддингиELMo [2].

Подобно экспериментам на англоязычных данных, мы также пробуем оценить взвешенную сумму предсказаний *FastText* и *ELMo*:

$$Score_S(w_1w_2) = (1-\alpha) * Score_{FT}(w_1w_2) + \alpha * Score_{ELMo}(w_1w_2) \quad (7)$$

FD-780				
	Kernel Regression		PLS Regression	
	Mean ($\|\rho\|$)	SD ($\|\rho\|$)	Mean ($\|\rho\|$)	SD ($\|\rho\|$)
CBOW-S (50)	0.43	0.06	0.42	0.06
α	MODEL-DP-S, CBOW vectors of dim. 50			
0.2	0.45	0.05	0.44	0.05
0.3	**0.45**	0.05	0.44	0.05
0.4	**0.45**	0.05	0.44	0.05
0.5	**0.45**	0.05	0.43	0.05
0.6	0.44	0.05	0.42	0.05
FD-1042				
	Kernel Regression		PLS Regression	
	Mean ($\|\rho\|$)	SD ($\|\rho\|$)	Mean ($\|\rho\|$)	SD ($\|\rho\|$)
CBOW-S (50)	0.42	0.05	0.42	0.05
α	MODEL-DP-S, CBOW vectors of dim. 50			
0.2	0.43	0.05	0.43	0.05
0.3	**0.43**	0.05	0.43	0.05
0.4	**0.43**	0.05	0.42	0.05
0.5	**0.43**	0.05	0.41	0.05
0.6	0.42	0.05	0.39	0.05

Табл. 5.1.1. Средняя корреляция Спирмэна и стандратное отклонение по 25 выборкам на экспериментах **FD-780** и **FD-1042**

5 Экспериментальная часть и обсуждение

5.1 Эксперименты на английских фразах

В **Таблице 5.1.1.** Представлены результаты эксперимента на датасете из 780 и 1042 примеров.

Как было сказано ранее, мы используем различные регрессионные модели для обучения на 75% датасета и тестирования на 25% датасета по 25 выборкам. Среди предложенных моделей, наилучшим образом показывают себя ядровая регрессия и регрессия PLS. В строке CBOW-S (50) представлены результаты для модели, использовавшей только предсказания на основе *Word2Vec*. В остальных случаях показаны результаты модели, комбинирующей предсказания *Poincaré* и *Word2Vec* с коэффициентом α. Можно пронаблюдать, что в обоих случаях базовый результат улучшается на 2 процентных пункта.

Метрика\Модель	FastText
cos (norm.)	**0.42**
L_∞ (avg.)	0.33
L_1 (avg.)	0.33
L_2 (avg.)	0.33

Табл. 5.2.1 Модуль корреляции Спирмэна (ρ) между ответами аннотаторов и метриками близости

5.2 Эксперименты на русских фразах

Таблица 5.2.1. показывает, насколько хорошо коррелируют расстояние между эмбеддингами слов и фразы по метрикам с оценками аннотаторов. Для L_1, L_2 и L_∞-метрик расстояние показывает отрицательную корреляцию с оценками. Это может быть объяснено природой векторов. Чем больше дистанция, тем "дальше"компаунд от его компонент в смысле семантики. Если смысл сочетания сильно отличается от смысла слов, она считается некомпозициональной. Чтобы результаты были сравнимы с предыдущими работами, мы предоставляем значения модуля корреляции.

Модель	ρ	Precision	Recall	F1	AUC-ROC
LSVC	**0.47**	0.37	**0.78**	0.48	0.83
MLP	**0.46**	0.32	**0.82**	0.44	**0.88**
DT	0.18	0.31	0.36	0.31	0.60
NB	**0.43**	**0.55**	0.52	**0.52**	0.71

Табл. 5.2.2 Результаты моделей машинного обучения на датасете с эмбеддингами FastText (метрики классификации представлены для класса 0).

В **Таблице 5.2.2.** показаны результаты различных моделей машинного обучения на датасете. Поскольку некомпозициональные компаунды представлены значительно хуже, и они представляют большую важность, метрики показаны для класса 0. Наибольшее значение по корреляции показывают метод опорных векторов и перцептрон, они же имеют более высокую полноту нулевого класса. Однако модель наивного Байеса показывает лучшую точность нулевого класса и значение F1. Что касается площади AUC, наибольшее значение показано перцептроном.

Классификаторы на эмбеддингах ELMo показывают более низкую корреляцию Спирмэна, но более высокий recall на классе 0 и более высокий AUC-ROC.

На сверточной нейронной сети получаем улучшение корреляции Спирм-

Модель	ρ	**Precision**	**Recall**	**F1**	**AUC-ROC**
LSVC	0.40	0.25	0.86	0.38	0.86
MLP	0.32	0.15	0.85	0.25	0.90
DT	0.11	0.27	0.27	0.26	0.55
NB	0.19	0.09	0.56	0.16	0.80

Табл. 5.2.3. Результаты моделей машинного обучения на датасете с эмбеддингами ELMo (метрики классификации представлены для класса 0).

Модель	ρ	**Precision**	**Recall**	**F1**	**AUC-ROC**
CNN	**0.53**	0.42	**0.83**	**0.53**	0.85

Табл. 5.2.4. Результаты сверточной нейронной сети на FastText(метрики классификации представлены для класса 0).

эна на 6 процентов.

Модель	ρ
LSVC	0.4960
MLP	0.4952
DT	0.2040
NB	0.4378
CNN	0.5389

Табл. 5.2.5. Модуль корреляции Спирмэна (ρ) для смешанного предсказания с α=0.25

Однакл, если смешать предсказания Word2Vec и ELMo, можно получить рост корреляции на 1 п.п.

6 Заключение

В этой работе был представлен новый метод оценки композициональности именных групп как на английском, так и на русском языке. В отличие от предыдущих работ, задача была рассмотрена с точки зрения предмета машинного обучения; было получено улучшение корреляции по сравнению с базовыми результатами. Кроме дистрибутивной информации по словам, была опробована также иерархическая, а именно — пуанкаре-эмебеддинги, содержащие информацию о гиперонимах и гипонимах элементов. В то время как сами по себе они не превосходят преыдущие исследования, в сочетании с *Word2Vec* наблюдается усиление корреляции на 2 процентных пункта.

Составлен первый датасет по композициональности именных групп для русского языка. Предоставлены базовые результаты по корреляции различных метрик с оценками аннотаторов. Далее они были значительно улучшены

с помощью построения классификаторов на эмбеддингах ELMo и Word2Vec. Было показано, что сочетание предсказаний на этих двух источниках также дает незначительное усиление качества.

В качестве дальнейшей работы планируется разработка методов для контекстуальной оценки композициональности. Элементы класса пока 2 были проигнорированы в классификации, в дальнейшем предлагается использовать контекст для разрешения в каждом конкретном случае.

Ныншений датасет имеет достаточно мало элементов, для применения более сложных алгоритмов необходимо его обогащение, в том числе за счет массовой разметки (Яндекс.Толока, Amazon Mechanical Turk). Возможно, наличие этой работы подтолкнет к созданию датасетов композициональных именных групп в некоторых других языках, в частности, восточно-славянских.

Список литературы

[1] Distributed representations of words and phrases and their compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen [и др.] // Advances in neural information processing systems. 2013. С. 3111–3119.

[2] Deep contextualized word representations / Matthew E. Peters, Mark Neumann, Mohit Iyyer [и др.] // CoRR. 2018. Т. abs/1802.05365. URL: http://arxiv.org/abs/1802.05365.

[3] В.В. Виноградов. Об основных типах фра- зеологических единиц в русском языке. 1947.

[4] Baldwin Timothy, Villavicencio Aline. Extracting the Unextractable: A Case Study on Verb-particles // Proceedings of CoNLL. 2002. С. 1–7. URL: https://www.aclweb.org/anthology/W02-2001.

[5] An Empirical Model of Multiword Expression Decomposability / Timothy Baldwin, Colin Bannard, Takaaki Tanaka [и др.] // Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18. 2003. С. 89–96. URL: https://doi.org/10.3115/1119282.1119294.

[6] McCarthy Diana, Keller Bill, Carroll John. Detecting a Continuum of Compositionality in Phrasal Verbs // Proceedings of the ACL 2003 Workshop

on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18. MWE '03. 2003. С. 73–80. URL: https://doi.org/10.3115/1119282.1119292.

[7] Venkatapathy Sriram, Joshi Aravind K. Measuring the Relative Compositionality of Verb-noun (V-N) Collocations by Integrating Features // Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. HLT '05. 2005. С. 899–906. URL: https://doi.org/10.3115/1220575.1220688.

[8] Reddy Siva, McCarthy Diana, Manandhar Suresh. An Empirical Study on Compositionality in Compound Nouns // Proceedings of the 5th International Joint Conference on Natural Language Processing. 2011. С. 210–218.

[9] How Naked is the Naked Truth? A Multilingual Lexicon of Nominal Compound Compositionality / Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio [и др.] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016. С. 156–161. URL: https://www.aclweb.org/anthology/P16-2026.

[10] Farahmand Meghdad, Smith Aaron, Nivre Joakim. A Multiword Expression Data Set: Annotating Non-Compositionality and Conventionalization for English Noun Compounds // Proceedings of the 11th Workshop on Multiword Expressions. 2015. С. 29–33. URL: https://www.aclweb.org/anthology/W15-0904.

[11] Gurrutxaga Antton, Alegria Iñaki. Combining Different Features of Idiomaticity for the Automatic Classification of Noun+Verb Expressions in Basque // Proceedings of the 9th Workshop on Multiword Expressions. 2013. С. 116–125. URL: https://www.aclweb.org/anthology/W13-1017.

[12] Roller Stephen, Schulte im Walde Sabine, Scheible Silke. The (Un)expected Effects of Applying Standard Cleansing Models to Human Ratings on Compositionality // Proceedings of the 9th Workshop on Multiword Expressions. 2013. С. 32–41. URL: https://www.aclweb.org/anthology/W13-1005.

[13] Yazdani Majid, Farahmand Meghdad, Henderson James. Learning semantic composition to detect non-compositionality of multiword expressions //

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015. C. 1733–1742.

[14] Lioma Christina, Hansen Niels Dalum. A Study of Metrics of Distance and Correlation Between Ranked Lists for Compositionality Detection // arXiv preprint arXiv:1703.03640v1. 2017.

[15] Segalovich Ilya. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine // MLMTA. 2003.

[16] Nickel Maximillian, Kiela Douwe. Poincaré Embeddings for Learning Hierarchical Representations // Advances in Neural Information Processing Systems 30 / под ред. I. Guyon, U. V. Luxburg, S. Bengio [и др.]. Curran Associates, Inc., 2017. C. 6338–6347. URL: http://papers.nips.cc/paper/7213-poincare-embeddings-for-learning-hierarchical-representations.pdf.

[17] TAXI at SemEval-2016 Task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling / Alexander Panchenko, Stefano Faralli, Eugen Ruppert [и др.] // Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). 2016. C. 1320–1327.

[18] Hearst Marti A. Automatic Acquisition of Hyponyms from Large Text Corpora // Proceedings of the 14th Conference on Computational Linguistics - Volume 2. COLING '92. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992. C. 539–545. URL: https://doi.org/10.3115/992133.992154.

[19] Platt John C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods // Advances in large margin classifiers. 1999. C. 61–74.

[20] Hinton G. E. Connectionist Learning Procedures // Artif. Intell. 1989. Т. 40, № 1-3. C. 185–234. URL: http://dx.doi.org/10.1016/0004-3702(89)90049-0.

[21] Classification and Regression Trees / L. Breiman, J. Friedman, R. Olshen [и др.]. Springer, 1984.

[22] Zhang Harry. The Optimality of Naive Bayes // Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004. T. 2. 2004. 01.