УДК 004.85

ОПЕРЕДЕЛЕНИЕ РЕГИОНА АВТОРА ПО ДАННЫМ ЖИВОГО ЖУРНАЛА

Морозов Е. В., Богданова Д. Н.

ООО «Аби ИнфоПоиск», Москва, Россия (127273, Москва, ул. Отрадная, 2Б.6), e-mail: eugene_m@abbyy.com

В настоящей работе представлен корпус записей русскоязычных блогов с информацией о местоположении автора, а также проведено исследование методов машинного обучения для автоматического определения региона автора. Для создания корпуса использовалась коллекция текстов блоговой платформы Живой Журнал (http://livejournal.com). Регионы авторов были приведены к единому виду, после чего из них были выбраны регионы с наибольшим количеством текстов. Корпус был очищен от выбросов – текстов, не представляющих интереса с точки зрения данного исследования. В данном исследовании были изучены различные наборы признаков, размеры обучающих коллекций и методы машинного обучения. Проведённые эксперименты показали, что большая часть текстов не содержит достаточно информации для определения региональной привязки, однако имеется существенная часть текстов, пригодных для региональной классификации.

Ключевые слова: геоклассификация, Живой Журнал, регион.

GEOGRAPHIC LOCATION PREDICTION IN BLOGS

Morozov E. V., Bogdanova D. N.

LLC «Abi InfoPoisk», Moscow, Russia (127273, 2B.6, Otradnaya, Moscow), e-mail: eugene_m@abbyy.com

This paper presents research on geographical lexical variation detection. We present a new corpus of Russian blogs labeled with geographical information. The corpus was extracted from LiveJournal (http://livejournal.com). Only those blogs that contained enough information about the author were used in the experiments. We have performed outlier detection, and thus, removed spam and other irrelevant data. We have studied various feature sets and performed classification based on Support Vector Machine and Naïve Bayes algorithms. The obtained results show that geographic location prediction is a hard task, and many of the blogs do not contain enough information to determine location of their authors, even though in certain cases accurate classification is possible.

Keywords: geographic location predicion, blogs, LiveJournal, region.

Введение

Изучение региональной специфики языка представляет большой интерес с точки зрения лингвистики и, кроме того, является важным звеном в задачах профилирования автора. Бурное развитие блогосферы в последние годы привело к появлению большого количества текстов с географической привязкой, что предоставило возможности для экспериментов в этой области, в частности с использованием алгоритмов машинного обучения. В данной работе проводится исследование русскоязычных текстов Живого Журнала. Каждый блог Живого Журнала имеет связанную с ним страницу – профиль автора, содержащую такую информацию об авторе, как возраст и местоположение. Для экспериментов использовались блоги, профили авторов которых содержали информацию о местоположении. Используемые алгоритмы были основаны на машинном обучении с учителем.

Ранее рядом авторов изучались вопросы, близкие к рассматриваемым в настоящей работе, на материале англоязычных блогов. В [5] изучалась геоклассификация текстов

блоговых платформ, но местоположение определялось не по языковым особенностям, а по IP-адресу, кроме того, в работе анализировалось всего 900 текстов. В [3] проводилась геоклассификация авторов записей Twitter, было показано, что при достаточном объёме сообщений (более ста) определение местоположения автора возможно с точностью более 51 %. В [8] представлен метод географической классификации сайтов, помимо используемого языка, определение местоположения, как и в [5], основано на таких характеристиках сайта, как домен и IP-адрес.

В данном исследовании такие характеристики, как информация о домене и IP-адрес, не используются, поскольку целью исследования является определение географической вариации языка автора, а не его местоположения. Другими словами, автор, обладающий языком определённого региона, должен классифицироваться как автор этого региона даже в том случае, когда территориально он находится в другом месте.

Целью настоящей работы является исследование географической вариации языка автора и возможности географической классификации большого количества (от 100 000 страниц и более) русскоязычных текстов на основе языковой специфики текстов, без использования информации об IP-адресе и домене.

Описание данных

Для исследований использовался корпус, составленный из русскоязычных текстов Живого Журнала. Живой Журнал представляет собой множество блогов, называемых журналами. Каждый журнал представлен набором страниц следующих трёх типов:

1. **Пост.** Страница представляет собой текст, составленный автором, и набор комментариев к нему, написанных пользователями Живого Журнала, других социальных сетей или анонимно.

- **Профиль**. Страница содержит информацию об авторе. Профиль содержит такие поля, как возраст, местоположение, места учебы и т. д. Также на этой странице автор может представить любую другую информацию в свободной форме. Кроме того, на данной странице содержатся ссылки на всех авторов, читающих данный журнал или читаемых данным автором, называемых друзьями пользователя.

2. **Главная страница журнала**. Страница содержит записи автора в хронологическом порядке, а также ссылки на страницы типа 1 и 2.

LiveJournal содержит как персональные журналы, все записи в которых создаются одним автором, так и коллективные журналы, называемые сообществами. В созданном нами экспериментальном корпусе минимальной текстовой единицей является запись, что позволяет идентифицировать авторов и включить в корпус не только тексты из персональных журналов, но и тексты сообществ.

Тексты журналов были получены с помощью поискового робота Apache Nutch [10]. Для выделения текста журнала и удаления html-обвязки был создан парсер страниц Живого Журнала. Для написания парсера были проанализированы типичные варианты оформления страниц и созданы правила на языке XPath, указывающие, какую часть html-страницы необходимо поместить в результирующий документ. Было экспериментально выяснено, что предложенные правила позволяют выделить текст для более чем 80 % русскоязычного содержимого Живого Журнала.

В рамках данного исследования использовался следующий алгоритм получения страниц:

1. Было задано стартовое множество страниц.
2. На каждом шаге случайным образом выбиралось некоторое множество подлежащих обработке страниц. Они обрабатывались, из них извлекались ссылки; ссылки сохранялись.

Главное отличие данного алгоритма от известного алгоритма PageRank [6] в том, что страницы выбирались случайным образом, а не в зависимости от количества ссылок на данную страницу, таким образом, в корпус вошли не только популярные журналы, но и те журналы, исходящих ссылок на которые было немного.

Помимо текстов записей, важная информация для формирования обучающего множества, содержится на страницах профилей (тип 2). В данной работе для формирования размеченного множества использовались сведения, указанные автором в графе «Местонахождение». Эксперименты основаны на предположении, что данные, указанные автором в этой графе, верны. Кроме того, предполагается, что автора можно считать носителем языковой специфики данного региона. Местоположение в рамках данного исследования – это наилучшее возможное приближение языковой вариации, и мы полагаем, что в большинстве случаев такое допущение правомерно, хотя требуются дополнительные исследования этого вопроса. Поле «местонахождение» заполняется в свободной форме, то есть автору предоставляется возможность указать, как город или посёлок, так и страну или регион. Например, возможно как такое написание «Российская Федерация, Московская область, Химки», так и просто «Российская Федерация» или «Химки». Кроме того, в названиях часто содержались опечатки, ошибки, а также встречались различные варианты написания названия одного и того же региона. Например, для Санкт-Петербурга встречались варианты: *Санкт-Петербург, Saint Petersburg, spb, SPb, Питер, Петербург* и другие. Всевозможные вариации названия одного и того же региона были приведены к единому виду. Далее было выполнено выделение крупных региональных единиц, язык которых предположительно может отличаться от других. Мелкие населенные пункты были присоединены к региональным центрам. Авторы, указавшие зарубежные регионы, как

правило, объединялись на уровне государства. Например, итоговое региональное деление содержало Москву, Нижегородскую область, Краснодарский край, Израиль, США. Для данного исследования не рассматривались авторы, принадлежащие иностранным государствам, за исключением Украины. Также была исключена Москва, так как, согласно данным исследований из других областей, большая доля проживающих в Москве родилась и выросла в других регионах, соответственно, может использовать лексику, специфичную для своих родных регионов. Из оставшихся регионов было отобрано 17 регионов с наибольшей долей русскоязычных текстов: Башкортостан, Челябинская область, Донецкая область, Киев, Омская область, Пермский край, Санкт-Петербург, Краснодарский край, Красноярский край, Московская область, Приморский край, Ростовская область, Самарская область, Саратовская область, Свердловская область, Татарстан. Всего отобрано примерно два миллиона сто тысяч текстов, считая комментарии, принадлежащие вышеописанным регионам. Из этих текстов только 184882 содержали более 300 слов.

Удаление выбросов

Эксперименты проводились на разных подмножествах полученных текстов Живого Журнала. В том числе для экспериментов было составлено тестовое подмножество, не содержащее выбросов. Удаление выбросов необходимо для очистки корпуса от спама – текстов, не представляющих лингвистического интереса и часто сформированных автоматически. Был составлен список из ста наиболее употребляемых слов во всем корпусе. В большинстве это были служебные слова и местоимения. Для каждого текста было подсчитано отношение содержащихся в нем слов из данного списка к общему количеству слов в тексте. Также для каждого текста подсчитана средняя длина предложений и общее количество токенов. На основании этих данных была проведена фильтрация коллекции. В первую очередь из коллекции были удалены тексты с аномально малой и аномально большой длиной предложений. Тексты с короткими предложениями, как правило, состояли из отдельных слов и восклицаний, а тексты с длинными предложениями представляли собой синтаксические не связанное перечисление различных объектов. Кроме того, из коллекции были удалены тексты с высокой концентрацией распространенных слов. Большая часть таких текстов представляла собой либо короткие наборы часто употребляемых слов, либо стихи и иные нетипичные экземпляры коллекции. Также выбросами считались тексты с низкой концентрацией распространенных слов, которые также часто оказывались стихами или спам-сообщениями. После удаления выбросов в коллекции осталось 1308870 текстов.

Отбор характеристик

Представление документов и выбор характеристик оказывает большое влияние на качество последующего машинного обучения. В случае текстовой классификации чаще всего

используются лексические характеристики – слова (униграммы), биграммы, триграммы и т.д. В рамках данной задачи использование таких характеристик целесообразно, поскольку задачей является определение лексической вариации языка. На представленной выше коллекции отбор характеристик осложняется большими объемами данных – общее количество лексических характеристик доходит до десятков тысяч. Всего было составлено четыре набора признаков. Для формирования характеристик первого набора использовался специально подготовленный лингвистами региональный словарь, основанный на данных словаря русских городов (http://community.lingvo.ru/goroda/). Используемый словарь состоял из 710 слов, специфических для отдельных русскоязычных регионов.

Для формирования следующих трёх наборов использовались т.н. ключевые слова. Для всего корпуса был составлен список слов вместе с частотой их использования. Для каждого региона был составлен аналогичный список на основе текстов, принадлежащих данному региону. После чего список ключевых слов для региона строился следующим образом: для очередного слова на основе количества вхождений в основной корпус и региональный подкорпус вычилялась LL-мера (log-likelyhood) следующим образом: пусть a – частота слова в первом корпусе, b – частота слова во втором корпусе, c – общее число слов в первом корпусе, d – общее число слов во втором корпусе. Тогда вычисляем $E_1=c(a+b)/(c+d)$, $E_2=d(a+b)/(c+d)$. Результирующим значением LL-меры будет $LL=2((a\ln(a/E_1))+(b\ln(b/E_2)))$. Подробное описание LL-меры можно найти в [7]. Список ключевых слов для каждого региона был отсортирован по значению LL-меры.

Для формирования второго набора признаков все слова вместе с их LL-мерами были объединены и отсортированы по убыванию LL-меры. Затем первые двенадцать тысяч были записаны в следующий файл. После удаления англоязычных слов, дубликатов осталось 10212 признаков.

Для третьего набора были выбраны по 800 первых ключевых слов для каждого региона. После их объединения и очистки осталось 9408 признаков, образовавших третий набор. Затем к признакам третьего набора были прибавлены слова из регионального словаря, составляющие первый набор. Получившийся набор признаков назовем четвертым набором.

Описание экспериментов

В качестве среды для машинного обучения использовались weka и liblinear ([4], [9]). Для экспериментов с первым набором признаков были отобраны тексты, содержащие более 300 слов. Только 4611 текстов из 184882 содержали слова из регионального словаря. Однако предварительные эксперименты показали, что многие из текстов, содержащих лексику словаря, содержат слова указанного региона. Например, около половины текстов, местоположением которых являлась Украина и содержащих слова из регионального словаря,

содержали слова как раз данного региона. Этот предварительный эксперимент показывает, что словарь не может быть использован как единственное средство классификации.

Со вторым и третьим набором признаков эксперименты проводились с помощью weka с использованием метода опорным векторов и Наивного Байесовского метода. Описание данных методов можно найти, например, в [1]. Также рассматривались только тексты, состоящие из не менее чем 300 слов. Эксперименты с алгоритмом SVM (используется библиотека LibSVM [2]) с полиномиальным ядром проводились на небольших подмножествах отобранных текстов. Кросс-валидация с разбиением на 10 частей показала точность в 15–20 %.

Эксперименты с Наивным Байесовским классификатором проводились и на всей отобранной коллекции, и на ее подмножествах. Данный классификатор показал несколько лучшие результаты, чем SVM. Это связано со слишком большой размерностью пространства признаков и малым количеством обучающих примеров для алгоритма SVM. Результаты работы Наивного Байесовского классификатора и SVM приведены на рисунке 1 (По оси Ox отмечена использованная доля коллекции, по Oy – процент правильных ответов на тестовой выборке).

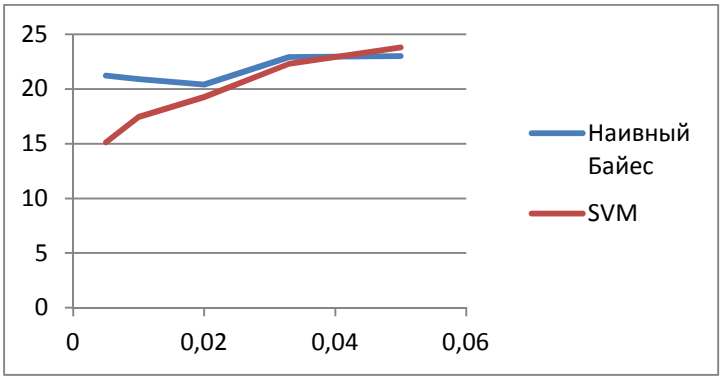

Рисунок 1. Результаты работы Наивного Байесовского классификатора и SVM

При обучении на всей коллекции Наивный Байес правильно классифицировал 24 % тестовой выборки. Таким образом, качество работы данного классификатора оказалось мало зависимым от размера тренировочных данных в данной задаче.

Для тестирования SVM на всей коллекции применялся пакет liblinear. Главное его достоинство – скорость работы, что дает возможность работать с большими объемами данных за счет отказа от использования нелинейных ядер в SVM. Liblinear тестировался на выборке, очищенной от выбросов. Результат для коллекции из текстов, содержащих не менее 20 слов, – 15 %. Для текстов длиной от 300 слов результат заметно лучше – 35 %. Это говорит о том, что короткие тексты практически невозможно привязать к конкретному региону, в то время как на больших текстах проявляются закономерности. Кроме того, были

проведены эксперименты с помощью liblinear на следующих размерах обучающей коллекции с более чем 300 словами: 1/16, 1/8, 1/4, 1/2 части обучающей коллекции. Результаты представлены на рисунке 2 (по оси Ox отмечена доля обучающей коллекции, по оси Oy – процент правильных ответов).

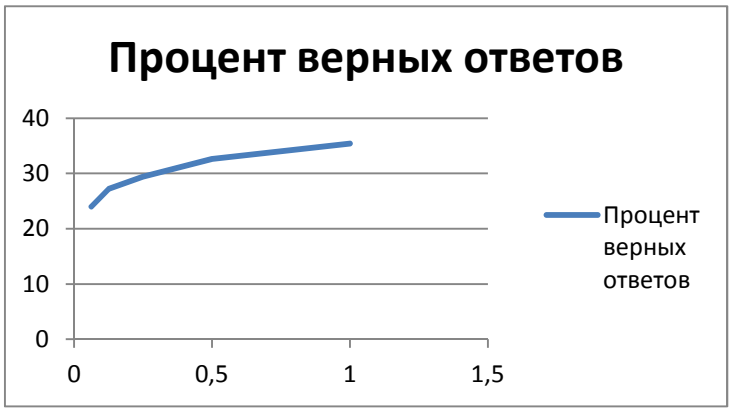

Рисунок 2. Результаты экспериментов, выполненных с помощью liblinear

Заключение

В рамках данного исследования был построен корпус записей Живого Журнала с географической привязкой. Эксперименты на созданном корпусе подтверждают трудность задачи географической классификации. При этом особые трудности для географической классификации представляют короткие тексты. Скорее всего, короткие тексты не содержат достаточно информации для определения географической вариации языка. С другой стороны, результаты классификации на коллекции, очищенной от выбросов и состоящей из текстов достаточной длины, говорят о наличии некоторых закономерностей. Возможно, лучшим решением для задачи геоклассификации является создание алгоритма, отказывающегося классифицировать большинство текстов, но показывающего высокую точность.

Авторы благодарят Сергея Александровича Шарова и Владимира Павловича Селегея за внимание к работе и полезные советы.
Работы проводятся при финансовой поддержке Министерства образования и науки Российской Федерации.

Список литературы

1. Bishop C. M. Pattern Recognition and Machine Learning // Springer, Series: Information Science and Statistics, 2006. – 740 p.
2. Chang C. C., Lin C. J. LIBSVM: a library for support vector machines // ACM Transactions on Intelligent Systems and Technology. – 2011. – P. 2-27.

3. Cheng Zhiyuan, Caverlee James, Lee Kyumin. You are where you tweet: a content-based approach to geo-locating twitter users // CIKM ’10 Proceedings of the 19th ACM international conference on Information and knowledge management. – 2010. – P. 759-768.
4. Fan R. E., K. W. Chang, C. J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification // Journal of Machine Learning Research. – 9(2008). – P. 1871-1874.
5. Fink Clay, Piatko Christine, Mayfield James, Finin Tim, Martineau Justin Geolocating Blogs From Their Textual Content // Working Notes of the AAAI Spring Symposium on Social Semantic Web: Where Web 2.0 meets Web 3.0 – 2009.
6. Page L., Brin S., Rejeev M., Terry W. The PageRank Citation Ranking: Bringing Order to the Web // Technical Report. Standford InfoLab,1999.
7. Rayson P., Garside R. Comparing corpora using frequency profiling // 38th annual meeting of the Association for Computational Linguistics. – 2000. – P. 1-6.
8. Volkov Alexey, Serdyukov Pavel. Unified Classification Model for Geotagging Websites // WWW2012 P. 625.
9. Weka 3: Data Mining Software in Java [Электронный ресурс]. – Режим доступа: http://www.cs.waikato.ac.nz/ml/weka/ (дата обращения 25.06.2013).
10. Welcome to Apache Nutch [Электронный ресурс]. – Режим доступа: http://nutch.apache.org/ (дата обращения 25.06.2013).

Рецензенты:

Беликов В.И., д.ф.н., доцент, кафедра теоретической и прикладной лингвистики филологического факультета Московского государственного университета имени М. В. Ломоносова, г. Москва.

Гриненко М.М., д.ф.-м.н., научный консультант, ООО «Аби ИнфоПоиск», г. Москва.