УДК 81

ЖАНРОВАЯ КЛАССИФИКАЦИЯ В ГЕНЕРАЛЬНОМ ИНТЕРНЕТ-КОРПУСЕ РУССКОГО ЯЗЫКА

Пиперски А.Ч.

Институт лингвистики ФГБОУ ВПО «Российский государственный гуманитарный университет», Москва, Россия (125993, Москва, Миусская пл., 6, корп. 2), e-mail: apiperski@gmail.com

Корпуса представляют собой важнейший инструмент современных лингвистических исследований. Для получения достоверных результатов исследователи, пользующиеся корпусами, должны обращать внимание на параметры метатекстовой разметки (информацию о социолингвистической, региональной, жанровой и т. п. принадлежности текста). В большинстве корпусов метатекстовые данные добавляются вручную, однако это невозможно при разработке больших корпусов, создаваемых на основе текстов из Интернета. Одним из таких корпусов является Генеральный интернет-корпус русского языка (ГИКРЯ), в котором применяются автоматические технологии метатекстовой разметки. В частности, предлагается новая схема жанровой разметки, при которой не выделяются априорные категории, а производится кластеризация на основе значений ряда переменных, выполняемая при помощи машинного обучения.

Ключевые слова: корпусная лингвистика, жанры, автоматическая разметка, кластеризация.

GENRE CLASSIFICATION IN THE GENERAL INTERNET CORPUS OF RUSSIAN

Piperski A.C.

Russian State University for the Humanities, Institute of Linguistics, Moscow, Russia (Miusskaya Sq. 6-2, 125993, Moscow), e-mail: apiperski@gmail.com

Corpora are indispensable research tool in present-day linguistics. If a scholar wants to achieve reliable results in a corpus-based study, he should take into account metadata, i.e. sociolinguistic, regional and genre-related properties of the texts included into the corpus. In most corpora metadata are added manually, which is not possible when constructing large Web-based corpora. Since the General Internet Corpus of Russian (GICR) is one of such corpora, it has to use automated metadata tagging. The developers of GICR propose a novel approach to genre classification without postulating any a priori categories. Machine learning algorithms are used to cluster texts based on automatically extractable features.

Keywords: corpus linguistics, genres, automated tagging, clustering.

Введение

Важным инструментом в работе современного лингвиста являются корпуса – информационно-справочные системы, основанные на собрании текстов на некотором языке (или языках) в электронной форме [4]. В российской лингвистике широко распространилось понимание того факта, что достоверным источником знаний о языке является не внутреннее ощущение исследователя, а материал, полученный из наблюдений над реальными текстами, представленными в корпусах [3].

Если придерживаться строгих методологических установок, никакой корпус не дает представления о языке в целом. Язык характеризуется широкой социальной, диахронической, географической, жанровой и т. п. вариативностью, а значит, любое суждение о языке в целом неизбежно будет некоторым приближением, полученным на материале определенной выборки. Поскольку состав генеральной совокупности текстов на языке не поддается определению (мы не можем достоверно сказать, какую долю всех текстов

на русском языке составляют юридические тексты, какую долю – тексты, написанные женщинами, и т. п.), не существует и методов, позволяющих достоверно установить, репрезентирует ли язык в целом выборка текстов, попавшая в корпус. Поэтому оказывается, что часто имеет смысл говорить не про язык как таковой, а про язык текстов того или иного жанра, про язык блогов, про язык текстов того или иного региона и т. п. Для того, чтобы осуществлять исследования подобного рода, необходимо сначала разработать корпуса, позволяющие их осуществлять. В таких корпусах тексты должны быть снабжены так называемой метатекстовой разметкой, то есть описанием их социолингвистических, жанровых, географических, хронологических и т. п. характеристик [7].

Проектом такого корпуса является Генеральный интернет корпус русского языка ([1], [2]), в котором планируется собрать репрезентативный срез русскоязычного Интернета и снабдить все тексты лингвистической и метатекстовой разметкой. Поскольку объем корпуса будет составлять не менее 20 миллиардов словоупотреблений, его ручное аннотирование не представляется возможным, и это приводит к необходимости создавать инструменты автоматической разметки. В настоящей статье речь пойдет в первую очередь об автоматической разметке по жанрам.

Классификация жанров

Систематику жанров ученые разрабатывают с глубокой древности [6]. Несмотря на это, единая классификация жанров до сих пор не выработана. В последние годы появился род новых подходов к автоматической классификации жанров, однако ни один из них так и не стал общепринятым [8].

Обычно создатели корпусов формируют список категорий, исходя из своих априорных представлений о том, как устроена система жанров. Важно, что система жанров, предлагаемая в корпусе, должна соответствовать представлениям о жанрах тех лингвистов, которые являются целевой аудиторией корпуса. В случае если разрыв между классификационной схемой корпуса и традиционными представлениями окажется слишком велик, есть риск, что многие лингвисты предпочтут игнорировать жанровое разнообразие языка и пользоваться корпусом как целым.

Разумеется, важным фактором является операциональность классификации: система жанров должна быть обозримой, чтобы ее могли использовать люди, размечающие корпус. Если постулировать небольшое количество жанров, тексты будут плохо вписываться в категории классификации, но если жанров окажется слишком много, разметчики потеряют способность в них ориентироваться. Так, в некоторых классификациях выделяется до 4000 различных жанров, среди которых есть, например, такие пункты, как диагноз, благодарственная песнь, брачный контракт, сценарий, план диеты и другие [5]. Разумеется,

для практического использования такой набор категорий неприменим, и в существующих корпусах набор категорий намного меньше. Например, в Национальном корпусе русского языка (www.ruscorpora.ru) для художественных текстов предлагается следующая система жанров:

1) нежанровая проза; 2) детектив, боевик; 3) детская; 4) историческая проза; 5) приключения; 6) фантастика; 7) любовная история; 8) юмор и сатира; 9) документальная проза; 10) драматургия; 11) перевод.

Кроме того, в этом корпусе имеется понятие «тип текста», и для художественных текстов выделяется такой набор категорий:

1) басня (прозаическая); 2) загадка; 3) записки; 4) легенда; 5) либретто; 6) миниатюра: анекдот; 7) миниатюра: шутка; 8) очерк; 9) письмо литературное; 10) повесть; 11) поэма (прозаическая); 11) притча; 12) пьеса; 13) рассказ; 14) роман; 15) сказка; 16) сказ; 17) сценарий; 18) цикл.

Неоднородность и непоследовательность такого деления очевидны, даже несмотря на то, что здесь мы имеем дело с подробно разработанной в литературоведении классификацией. Так, в число прозаических поэм (по состоянию на 01.06.2013) попадают такие тексты:

Венедикт Ерофеев. Москва-Петушки (1970)

А. С. Макаренко. Педагогическая поэма. (1933–1935)

Максим Горький. Человек (1903)

Максим Горький. Двадцать шесть и одна (1899)

А. Н. Радищев. Ангел тьмы (1780–1802)

С. Н. Сергеев-Ценский. Живая вода (1927)

С. Н. Сергеев-Ценский. Недра (1912)

С. Н. Сергеев-Ценский. Неторопливое солнце (1911)

С. Н. Сергеев-Ценский. Молчальники (1905)

В их число не вошло произведение Н. В. Гоголя «Мертвые души», которое классифицируется как роман. Однако очевидно, что нет никаких оснований считать произведение В. Ерофеева «Москва-Петушки» поэмой, а «Мертвые души» – нет: и то и другое – прозаические тексты, которые, однако, были охарактеризованы как поэмы самими авторами.

Если же перейти в область нехудожественных текстов, проблема становится еще более сложной, поскольку там нет общепринятых систем жанров. Кроме того, ясно, что для выделения подобных жанров необходима ручная разметка, которую можно обеспечить в

масштабах Национального корпуса русского языка, но невозможно – в масштабах Генерального интернет-корпуса русского языка.

В Генеральном интернет-корпусе русского языка используется не вполне традиционный подход к жанровой классификации. Вместо замкнутой классификации, основанной на априорных категориях, предлагается использовать альтернативный подход: жанровые категории выделяются апостериорно на основе сходства между собой текстов, входящих в корпус.

На первом этапе асессоры размечают достаточно большой обучающий корпус текстов, отобранных случайным образом, характеризуя каждый из них по ряду содержательных параметров. Список этих параметров и система оценки были разработаны специально для Генерального интернет-корпуса русского языка С. А. Шаровым. Для каждого текста асессоры должны дать ответ на 15 вопросов:

A1. До какой степени текст стремится побудить читателя поддержать какую-либо точку зрения (или отказаться от существующей)?

A2. До какой степени текст отражает точку зрения организации?

A3. До какой степени текст выражает чувства или эмоции автора?

A4. До какой степени текст говорит о вымышленных персонажах/реалиях?

A5. До какой степени текст предназначен для развлечения читателя?

A6. До какой степени текст написан в неформальном стиле, например, с использованием просторечия или сленга?

A7. До какой степени текст стремится обучить пользователя делать что-либо?

A8. До какой степени текст является информационным сообщением или похож на него в виде, который может появиться в газете?

A9. До какой степени текст носит юридический характер?

A10. До какой степени текст отражает устную речь?

A11. До какой степени текст написан от первого лица?

A12. До какой степени текст стремится продвигать коммерческий продукт или услугу?

A13. До какой степени текст стремится пропагандировать политическую партию, религиозные взгляды или другие группы с некоммерческими целями?

A14. До какой степени текст принадлежит к числу научных или технических текстов?

A15. До какой степени понимание текста требует знаний специалиста?

На каждый из этих вопросов асессор выбирает один из ответов на четырехчастной шкале:

0: ни в какой степени;

1: слегка;

2: частично;

3: большей частью.

Признаки, описанные выше, позволяют описывать самые разные тексты: они не предназначены специально для того или иного рода текстов и могут использоваться для большинства текстов, автоматически скачиваемых из Интернета.

Теоретически такая система признаков позволяет различить 4^{15} = 1073741824 жанра. В ходе пилотного эксперимента, проводившегося при помощи двух асессоров, было проанализировано, насколько разнообразными оказываются приписываемые текстам наборы признаков. По результатам разметки 1-го аннотатора 113 текстов были распределены на 86 категорий, а 2-й аннотатор распределил 111 других текстов по 67 категориям. В общей сложности 224 текста были распределены по 147 категориям (6 категорий у обоих асессоров оказались общими).

Разумеется, в таком разнообразии, которое получается при учете всех характеристик текста по отдельности, нет практической необходимости, и суть дальнейшей обработки сводится к тому, чтобы сгруппировать тексты со схожими характеристиками и тем самым построить апостериорную схему жанровой классификации.

На втором этапе из текстов обучающего корпуса извлекаются характеристики, легко получаемые автоматически, например, средняя длина предложения, количество предложений, употребительность некоторых наиболее частотных слов русского языка и т. п. Эти характеристики сопоставляются с содержательными характеристиками, предложенными асессорами, и при помощи статистических методов выявляются корреляции между формальными и содержательными признаками. На третьем этапе производится кластеризация текстов на основе алгоритмов машинного обучения, и таким образом обеспечивается распределение текстов по жанровым группам на основе формальных характеристик. Четвертый этап работы может включать в себя ручную обработку тестового подкорпуса с целью установить соответствие получившихся жанровых групп характеристикам традиционно выделяемых жанров.

Предлагаемый метод машинного обучения позволит осуществить жанровую классификацию всех текстов, входящих в Генеральный интернет-корпус русского языка. Разумеется, необходимо понимать, что при этом не удастся обеспечить стопроцентную точность разметки, однако она будет достаточно высокой для того, чтобы выполнять на материале корпуса лингвистические исследования, принимающие во внимание жанровую вариативность текстов, написанных на русском языке.

Работы проводятся при финансовой поддержке Министерства образования и науки Российской Федерации в рамках выполнения ГК 07.514.11.4142 по теме «Разработка методов автоматического и полуавтоматического создания корпуса и подкорпусов современного русского языка на основе русскоязычного Интернета» и программы стратегического развития РГГУ.

Список литературы

1. Беликов В. И., Селегей В. П., Шаров С. А. 2012. Пролегомены к проекту Генерального интернет-корпуса русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая – 3 июня 2012 г.). Вып. 11 (18). – М.: Изд-во РГГУ, 2012. – С. 37–50.

2. Беликов В. И., Копылов Н. Ю., Пиперски А. Ч., Селегей В. П., Шаров С. А. Корпус как язык: от масштабируемости к дифференциальной полноте // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая – 2 июня 2013 г.). – Вып. 12 (19). – М.: Изд-во РГГУ, 2013. – С. 84–95.

3. Плунгян В. А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. – 2008. – № 2 (16). – С. 7–20.

4. Что такое Корпус?: [Электронный документ]. – (http://ruscorpora.ru/corpora-intro.html). Проверено 01.06.2013.

5. Adamzik, K. Textsorten — Texttypologie. Eine kommentierte Bibliographie. – Münster: Nodus, 1995. – 301 p.

6. Corbett J. Genre and Genre Analysis // Encyclopedia of Language and Linguistics. Ed. by K. Brown. – Amsterdam, Boston: Elsevier, 2006. – P. 26-32.

7. McEnery T., Hardie A. Corpus Linguistics. – Cambridge: Cambridge University Press, 2011. – xv, 294 p.

8. Mehler, A., Sharoff, S., Santini, M. (eds.). Genres on the Web: computational models and empirical studies. – New York: Springer, 2010. – xiv, 362 p.

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.