NLPub: каталог и сообщество русских лингвистических ресурсов

© Дмитрий Усталов
Институт математики и механики им. Н.Н. Красовского УрО РАН
Екатеринбург
dau@imm.uran.ru

Аннотация

Разрозненность сведений о существующих инструментах и ресурсах для автоматической обработки русского языка является большой проблемой, сильно затрудняющей быстрый старт научных и практических работ, тормозя развитие всего направления. Наличие специализированного каталога лингвистических ресурсов позволит решить эту проблему хотя бы частично. В данной работе представлен каталог и сообщество NLPub, проведено сравнение с аналогичными проектами, описан используемый подход к сбору и представлению данных, продемонстрирована классификация разделов, кратко изложен опыт, полученный с момента основания проекта, и обозначены планы на ближайшее будущее.

1 Введение

Словари и тезаурусы, корпусы текстов и банки данных, а также другие информационные ресурсы, имеют огромную ценность в области обработки естественного языка. Это обусловлено спецификой фундаментальных и прикладных задач компьютерной лингвистики, нередко решаемых при помощи разнообразных статистических методов.

За последние годы популярность технологий автоматической обработки естественного языка заметно выросла благодаря таким продуктам, как Apple Siri, Wolfram|Alpha, Google Voice, и др. Возник закономерный общественный интерес, однако разрозненность русскоязычных лингвистических ресурсов затрудняет быстрый старт новых проектов в данной области.

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

Несмотря на ценность и очевидную как научную, так и коммерческую значимость исследований и разработок в области обработки естественного языка, сегодня наблюдаются следующие проблемы:

• отсутствие доступного качественного инструментария и вспомогательных утилит для обработки текста, для распознавания речи, и т.д.;

• нехватка доступных информационных ресурсов: машиночитаемых словарей, тезаурусов, размеченных корпусов текстов, банков данных;

• дефицит сведений об экспертах, тематических мероприятиях и образовательных программах в регионах.

Указанные проблемы делают особенно актуальной задачу сбора, систематизации и распространения сведений о доступных средствах и ресурсах для обработки русского языка.

Цель проекта NLPub[1] заключается в предоставлении на *некоммерческой* основе каталога электронных материалов, направленного на удовлетворение информационных потребностей пользователей, исследователей и разработчиков в области компьютерной лингвистики. Проект NLPub появился и развивается за счет личных средств автора и не имеет аффилированности со сторонними организациями.

2 Аналогичные работы

Среди подобных русскоязычных ресурсов можно отметить [1]:

• *Портал знаний о компьютерной лингвистике*[2], созданный в Институте систем информатики им. А.П. Ершова СО РАН, г. Новосибирск;

• *Лингвистика в России: ресурсы для исследователей*[3], созданный в Московском государственном университете им. М.В. Ломоносова, г. Москва;

• *Каталог лингвистических программ и ресурсов в Сети*[4], созданный в Русской виртуальной библиотеке, г. Москва;

• *Математическая и компьютерная лингвистика*[5], созданный в Санкт-Петербургском государственном университете, далее — *mathlingvo*.

2.1 «Портал знаний о компьютерной лингвистике»

Портал знаний по компьютерной лингвистике существует с 2006 г. и призван обеспечить систематизацию и интеграцию знаний и информационных ресурсов по компьютерной лингвистике в единое информационное пространство, а также содержательный доступ к интегрированным знаниям и ресурсам.

На портале представлены знания об основных разделах компьютерной лингвистики, о ее предмете и объектах исследования, используемых в ней моделях и методах, разработанных в рамках компьютерной лингвистики технологиях, системах, программных продуктах и лингвистических ресурсах (словарях, корпусах и лингвистических баз данных), а также информация об ученых, сообществах, организациях, включенных в процесс исследования по компьютерной лингвистики и о выполняемых проектах в этой области.

По всей видимости, развитие портала остановилось в 2012 г.

2.2 «Лингвистика в России: ресурсы для исследователей»

Научно-образовательный портал «Лингвистика в России: ресурсы для исследователей» создан также в 2006 г. по инициативе Научно-исследовательского вычислительного центра МГУ им. М.В. Ломоносова и Казанского государственного университета им. В.И. Ульянова-Ленина и имеет раздел, посвященный компьютерной лингвистике.

Задачей портала является создание инфраструктуры для поддержки сообществ исследователей и преподавателей для информирования и открытого обсуждения научных и образовательных задач российской лингвистики, интеграция лингвистического сообщества Российской Федерации. На портале собран каталог ссылок на различные российские проекты в области компьютерной лингвистики.

По всей видимости, развитие портала остановилось в 2007 г.

2.3 «Каталог лингвистических программ и ресурсов в Сети»

Данный каталог включает в себя описание программ, связанных с анализом текстов и вычислительной лингвистикой, а также соответствующих ресурсов, доступных в Интернете.

Упор при составлении каталога делался на бесплатные программы, доступные для загрузки. Однако также описаны некоторые сетевые и коммерческие версии программ. Тематически каталог разбит на следующие разделы: программы анализа и лингвистической обработки текстов; программы преобразования текстов; психолингвистические программы; генераторы текстов и «говорящие» программы; системы обработки естественного языка; коллекции ресурсов; словари и тезаурусы.

Развитием каталога занимается его единственный составитель, внося достаточно редкие дополнения, правки и изменения. Последнее обновление каталога зафиксировано в 2013 г.

2.4 «Математическая и компьютерная лингвистика»

mathlingvo — проект кафедры информационных систем в искусстве и гуманитарных науках Санкт-Петербургского государственного университета, созданный в начале 2012 г. и посвященный математической и компьютерной лингвистике в России.

Проект представляет собой коллективный блог под руководством представителей кафедры, в котором уделено внимание перечням тематических конференций, периодических изданий, вакансиям. Также является представительством различных инициатив, таких как OpenCorpora[6].

Лента новостей *mathlingvo* обновляется регулярно и поддерживает добавление новых записей от любого участника на условиях предварительной модерации, однако проект является в большей степени новостным ресурсом и не предоставляет собой каталог как таковой.

3 NLPub: каталог и сообщество

NLPub — это каталог лингвистических ресурсов для обработки русского языка, основанный на принципах краудсорсинга. День рождения проекта отмечается первого октября 2012 г., когда NLPub был представлен широкой общественности на «Хабрахабре» [2].

Каталог. Каталог построен на базе MediaWiki — программного обеспечения, лежащего в основе «Википедии» и «Викисловаря» (рис. 1). Основное отличие NLPub от аналогичных ресурсов, заключается в открытости: любой желающий может внести свои изменения по хорошо известным принципам «Википедии». Благодаря открытости и децентрализованности, материалы NLPub поддерживаются в актуальном, корректном и доступном состоянии с меньшими трудозатратами и большей заинтересованностью участников. Прототипом каталога послужил проект ACLWiki[7], созданный Ассоциацией по компьютерной лингвистике.

Сообщество. Важно отметить, что NLPub — это не только краудсорсинговый каталог лингвистических ресурсов, но и сообщество, представленное вокруг этого каталога, вопрос-ответного сервиса NLPub Q&A[8] на базе открытого движка Discourse, и Twitter-аккаунта @nlpub. Также на NLPub расположена и поддерживается документация проекта создания открытого электронного тезауруса русского языка Yet Another RussNet[9].

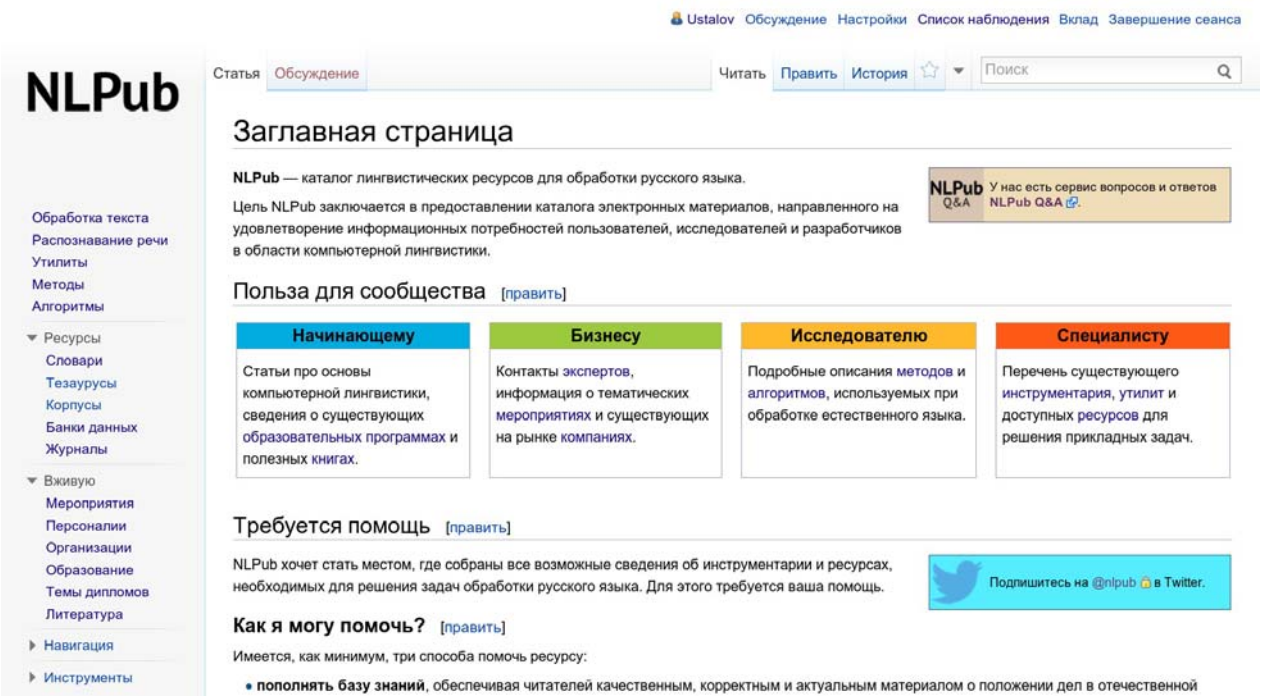

Рис. 1: Главная страница http://nlpub.ru/

4 Организация каталога

Каталог организован преимущественно в табличном виде и, в отличие от онтологического подхода [3], представляет собой квазиструктурированные данные в формате более привычной вики-разметки. Это упрощает пополнение и улучшение каталога со стороны человека. Таблицы содержат наиболее ценную информацию об отдельно взятом объекте. Например, для программного продукта в таблице приводится информация о кратком назначении, поддерживаемых языках и условиях использования, а для организации — год основания и ключевые лингвистические продукты.

Инструменты и утилиты. Различные инструменты обработки естественного языка (более 140 наименований), распознавания речи (более 20 наименований), утилиты для работы с языковыми моделями и обработки банков данных. Для некоторых инструментов существуют выделенные страницы с подробным описанием и инструкцией по применению. Такими инструментами являются, в частности, Greeb и TreeTagger.

Ресурсы. Под *ресурсом* понимаются данные и их производные, используемые в процессе обработки естественного языка: корпусы текстов (более 5 наименований), тезаурусы и словари (более 20 наименований), банки данных. Для некоторых ресурсов существуют выделенные страницы с подробным описанием и перечнем особенностей. Такими ресурсами являются, в частности, словарь Абрамова и YARN.

Методы и алгоритмы. Небольшое собрание достаточно важных методов и алгоритмов обработки естественного языка, записанное в виде псевдокода с кратким описанием особенностей и характеристик. Для некоторых алгоритмов существуют выделенные страницы, например про алгоритм удаленной интерполяции и об алгоритме Витерби.

Образование. Перечень тематических кафедр, вузов, курсов и программ переподготовки, полезных как начинающим, так и опытным исследователям и разработчикам в области обработки естественного языка.

Мероприятия. Список тематических мероприятий и конференций, посвященных обработке естественного языка и компьютерной лингвистике, где можно представить и обсудить результаты своей работы. Существуют выделенные страницы для ряда конференций, например для конференции АИСТ.

Организации. Раздел, полезный при поиске работы и при анализе российского рынка решений по обработке естественного языка. Включает в себя достаточно полный список основных игроков на отечественном рынке NLP-продуктов.

Литература. Список литературы, полезной для изучения и закрепления знаний об обработке естественного языка и компьютерной лингвистике. Включает ссылки как на учебные пособия, так и на методические указания.

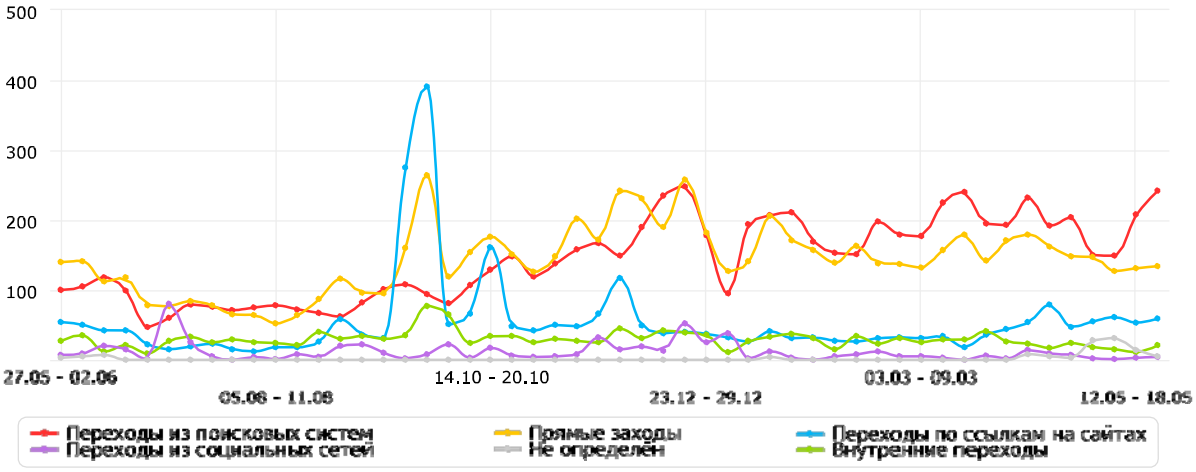

Рис. 2: Данные «Яндекс.Метрики» о посещаемости NLPub с 27 мая 2013 г. по 25 мая 2014 г.

Эксперты. Экспериментальный раздел, в котором любой желающий может указать область своей экспертизы и контактную информацию для выполнения какой-либо совместной работы или консультирования.

Темы дипломов. Экспериментальный раздел, в котором любой желающий может указать проблемную область, достойную разработки в рамках студенческой или кандидатской работы, и оставить свои координаты для связи.

5 Полученный опыт

Первые месяцы существования NLPub были сопряжены с борьбой против активных спам-ботов, специализирующихся на проектах, основанных на MediaWiki. Проблему удалось решить полностью благодаря одновременному принятию трех мер: введению капчи на основе reCAPTCHA при создании учетной записи, подключению черного списка спамерских IP-адресов, а также обязательным подтверждением адреса электронной почты для получения участником возможности вносить правки в статьи.

Данные «Яндекс.Метрики[10]» доступны публично и свидетельствуют о постепенном росте посещаемости NLPub за прошедший год (рис. 2). Это связано с тем, что по мере создания новых страниц и внесения новых сведений страницы становятся более ценными как с точки зрения читателей, так и с точки зрения поисковых машин.

Более высокие позиции в поисковой выдаче способствуют привлечению новых пользователей. Тем не менее, на текущий момент можно считать активность пользователей *эфемерной*, то есть человек попадает на NLPub во время поиска ответа на свой вопрос при помощи поисковых систем. Это свидетельствует о том, что база постоянных читателей и авторов недостаточно велика: упоминание ресурса в популярных блогах или сайтах отражается в статистике как резкий скачок вверх.

В настоящий момент сообщество находится на достаточно ранней стадии своего развития, однако уже сегодня на NLPub Q&A можно получить ответы на достаточно острые и нетривиальные тематические вопросы.

6 Заключение

Анализ поисковых запросов и опрос аудитории NLPub показывает заинтересованность в отдельных статьях, посвященных конкретным инструментам, методам и алгоритмам. Эта информация обобщена на специальной странице http://nlpub.ru/TODO. Выделяется три направления предстоящей работы:

- общие статьи об основных разделах автоматической обработки естественного языка: графематический, морфологический, синтаксический анализ, информационный поиск, сходство документов, машинный перевод, извлечение ключевых слов, автоматическое реферирование, анализ тональности, и др.;
- статьи о популярных моделях, методах и алгоритмах: векторные модели (tf-idf, «мешок слов», косинусная мера близости), теоретико-графовые модели, n-граммные модели, общие методы алгоритмического обучения, используемые в лингвистике (перцептрон, наивный Байесовский классификатор, EM-алгоритм), и др.;
- обучающие статьи о важном или слабо документированном программном обеспечении: «Томита-парсер», FreeLing, Stanford NLP, MaltParser, NLTK, и др.

На сегодняшний день можно отметить два основных недостатка ресурса. Во-первых, слабая заполненность некоторых разделов, таких как «Персоналии» и «Литература». Это вызвано достаточно небольшим возрастом NLPub и предполагается, что эта проблема решится путем органического роста проекта. Во-вторых, отсутствие связей между разными разделами каталога усложняет навигацию. Решение этой проблемы состоит в добавлении соответствующих внутренних

ссылок и предоставлении наглядной карты сайта на одной из главных страниц ресурса.

Повышение охвата пользователей и снижение эфемерности их активности можно выполнить путём интеграции с ресурсом *mathlingvo* для автоматической публикации сводок новостей с указанием соответствующих ссылок.

В отдаленной перспективе было бы интересно преобразовать каталог NLPub в семантическую вики для предоставления машиночитаемых данных с одновременным сохранением удобства внесения правок и дополнений в материалы проекта.

Благодарности. Автор выражает огромную благодарность всем пользователям NLPub, принявшим участие в работе над материалами проекта.

Литература

[1] Д. А. Усталов. Каталоги лингвистических ресурсов: состояние и перспективы // Молодой ученый. — 2012. — Т. 1, №12 (47). — С. 148–152.

[2] Д. А. Усталов. NLPub — каталог лингвистических решений. http://habrahabr.ru/post/152429/

[3] Ю. А. Загорулько и др. Подход к построению предметной онтологии для портала знаний по компьютерной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог». — 2006. — С. 148–151.

Примечания

[1] http://nlpub.ru/
[2] http://uniserv.iis.nsk.su/cl/
[3] http://uisrussia.msu.ru/linguist/_B_comput_ling.jsp
[4] http://www.rvb.ru/soft/catalogue/catalogue.html
[5] http://mathlingvo.ru/
[6] http://opencorpora.org/
[7] http://aclweb.org/aclwiki/
[8] http://qa.nlpub.ru/
[9] http://russianword.net/
[10] https://metrika.yandex.ru/stat/?counter_id=17329045

NLPub: a Catalogue and a Community for Russian Linguistic Resources

Dmitry Ustalov

The lack of coordination in the information on existing tools and resources for Russian language processing has become a significant problem. Such a problem complicates both research and practical applications thwarting with the progress of the whole field. A specialized catalogue for linguistic resources may assist one in getting this problem solved. In this survey NLPuba catalogue and a community for Russian linguistic resources is presented and compared with its analogs. Its data gathering and representation approaches are also described and the merotomy is demonstrated. The experience obtained since the project start is outlined and future work directions are stated.