Инженерный вестник Дона, №1 (2018)
ivdon.ru/ru/magazine/archive/n1y2018/4791

Автоматизация процесса коллективного построения лингвистических ресурсов

Д.А. Усталов[1,2], *А.Ю. Берсенёв*[1,2], *Ю.А. Киселёв*[2]

[1]*Институт математики и механики им. Н.Н. Красовского Уральского отделения Российской академии наук, Екатеринбург*
[2]*Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, Екатеринбург*

Аннотация: Статья посвящена построению и расширению лингвистических ресурсов, таких как словари и тезаурусы, при помощи краудсорсинга на основе выполнения микрозадач. Описаны подходы к априорной оценке сложности микрозадач, оценке производительности участника, адаптивному назначению микрозадач участникам. Представлен комплекс программ, спроектированный на основе трёхзвенной архитектуры, позволяющий использовать краудсорсинг для выполнения микрозадач. Описанные подходы доступны в составе представленного комплекса программ.
Ключевые слова: лингвистический ресурс, языковой ресурс, краудсорсинг, синсет, связь, микрозадача, вычислительная семантика, обработка естественного языка, человеко-машинная система, анализ данных, программное обеспечение.

Введение

Онтологии и другие формы представления знаний широко используются при решении задач интеллектуализации обработки информации [1]. При автоматической обработке текста на естественном языке применяются лингвистические (или *языковые*) ресурсы — специальные наборы данных, представляющие знания о языке. В отличие от большинства западноевропейских языков, для которых доступны высококачественные лингвистические ресурсы различного типа, подобные наборы данных для русского языка в настоящее время активно развиваются. Объём работы, необходимый для создания языковых ресурсов высок, поэтому исследователи широко используют краудсорсинг для эффективного сбора данных [2]. Краудсорсинг зарекомендовал себя в задачах разметки корпуса текстов силами волонтёров [3], коллективного обсуждения рационализаторских предложений [4], а также при построении лексико-семантических ресурсов [5,6]. В общем виде, краудсорсинг на основе

выполнения микрозадач состоит из трёх основных этапов: назначение задания участнику, получение ответа от участника, агрегация ответов после завершения работы всех участников. Каждое задание выполняется несколькими разными участниками для обеспечения надёжности результата. В данной статье представлен комплекс подходов к развитию языковых ресурсов при помощи краудсорсинга на основе выполнения микрозадач.

Оценка сложности заданий

Эксперименты на наборах данных OpenCorpora [3], RDT [7], LRWC [8], а также при построении семантических классов [9] показывают, что априорная оценка сложности заданий при построении лингвистических ресурсов является трудной задачей из-за большого количества субъективных факторов, не имеющих непосредственного отношения к лингвистическим особенностям размечаемых данных. Это свидетельствует о том, что построение универсальной меры априорной сложности заданий является трудной задачей. В свою очередь, целесообразно использовать следующий подход. Поскольку каждое задание выполняется несколькими разными участниками, то при получении каждого ответа следует измерять какую-либо меру согласованности или несогласованности ответов, например, информационную энтропию. Задания со значением меры согласованности ниже порогового размечать до тех пор, пока не будет достигнуто либо ожидаемое пороговое значение, либо ограничение по бюджету. Несмотря на то, что такой подход не предоставляет строгих гарантий аккуратности, возникает возможность фильтрации полученных ответов на основе производительности участников при агрегации результатов [9]. Таким образом, в качестве меры сложности заданий возможно использование таких мер, как информационная энтропия для ответов, принадлежащих номинальной шкале, и стандартное отклонение для ответов, принадлежащих шкале отношений.

Оценка производительности участника

Вопрос автоматической оценки производительности участника хорошо представлен в литературе [2]; существуют методы, гарантирующие оптимальный результат в определённых частных случаях. Наиболее популярным на практике подходом является метод Давида-Скина, предложенный в 1979 г. для задач медицинской диагностики задолго до возникновения краудсорсинга. Данный метод основан на EM-алгоритме и строит матрицы ошибок участников, на основании которой можно вычислить аккуратность участников. Другим популярным методом является Wawa (англ. *worker agreement with aggregate*), определяющий производительность участника как отношение количества ответов этого участника, совпавших с ответами других участников, к количеству всех ответов данного участника.

Назначение заданий

На основе подходов к оценке производительности участника и назначения заданий возможно построить функцию адаптивного назначения заданий. Пусть задано множество заданий T и участник w, которому необходимо назначить задание; пусть также заданы $d(t)$ — мера сложности некоторого задания t, $p(w)$ — мера производительности участника w. В таком случае возможно установить связь между квалификацией участника $p(w)$ и сложностью задания $d(t)$, сила которой характеризуется числовым параметром k в виде следующего правила подбора назначенного задания t^*:

$$t^* = \arg\max_{t \in T} k \cdot d(t) \cdot p(w).$$

Комплекс программ

В практических задачах необходимо экспериментировать с различными конфигурациями методов назначения заданий, оценки участников, агрегации ответов. Инженерная реализация даже тривиальных методов требует времени на программирование и тестирование. Многие

популярные подходы, в том числе описанные в данной работе, реализованы в комплексе программ Mechanical Tsar. Данный комплекс программ основан на трёхзвенной архитектуре: «*интерфейс – движок* (*сервис*) *– база данных*», изображённой на рис. 1. Такая организация системы позволяет использовать несколько различных интерфейсов для одних и тех же данных одновременно. Например, как Web-интерфейс, так и Telegram-бота.

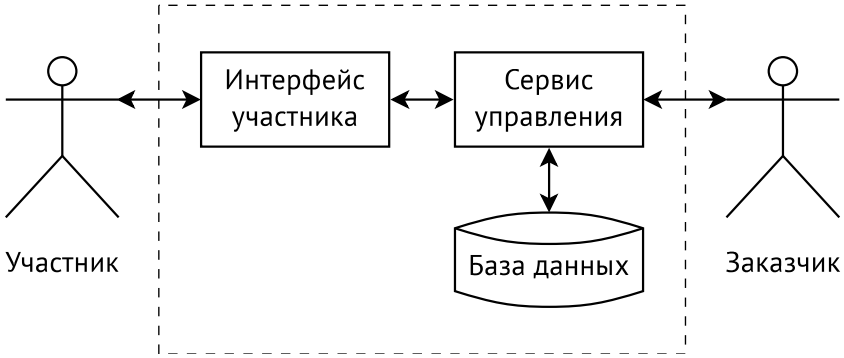

Рис. 1. – Архитектура комплекса программ Mechanical Tsar

Аудит исходного кода разработанного комплекса программ не выявил проблем с обработкой внешних запросов, поскольку как интерфейс участника, так и сервис управления реализованы с использованием абстракций, осуществляющих должную фильтрацию как HTTP-запросов, так и SQL-запросов. В свою очередь, следует отметить следующие риски при эксплуатации:

1. запуск сервиса управления в открытой недоверенной сети небезопасен, поскольку программный и административный интерфейс сервиса доступен по протоколу HTTP или HTTPS без аутентификации;

2. отсутствие принудительного использования шифрования и протокола HTTPS в интерфейсе участника допускает перехват трафика в общедоступных сетях;

3. идентификация участника по неаккуратно составленной неуникальной метке допускает получение заданий и отправку ответов от чужого имени.

Построение лингвистических ресурсов

Описанные в данной работе подходы успешно использовались при построении двух наборов данных для русского языка и одного набора данных для английского языка:

4. При построении набора данных RDT (англ. *Russian Distributional Thesaurus* — русский дистрибутивный тезаурус) производился сбор суждений людей о попарной семантической близости русских слов [7].

5. Набор данных LRWC (англ. *Lexical Relations from the Wisdom of the Crowd* — «мудрость толпы о лексических связях») представляет суждения людей об осмысленности иерархических связей между русскими словами на примере гипонимов и гиперонимов [8].

6. При оценке метода построения семантических классов для английского языка проверялось соответствие слов автоматически определённому семантическому классу и обратное соответствие [9].

Упомянутые языковые ресурсы доступны для использования на условиях открытой лицензии Creative Commons Attribution-ShareAlike.

Заключение

В статье были продемонстрированы подходы к повышению эффективности коллективного построения лингвистических ресурсов при помощи краудсорсинга на основе выполнения микрозадач. Исходный код представленного в данной работе комплекса программ доступен на сайте GitHub: https://github.com/mtsar. В качестве направлений дальнейших исследований следует выделить снижение входного барьера для новых участников краудсорсинговых проектов и разработку подходов к повышению заинтересованности имеющихся участников [10].

Благодарность. Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 16-37-00354 мол_а.

Литература

1. Ефремова О.А., Абдуллина Й.И., Юферова А.В. Интеллектуализация программного обеспечения по обработке пространственных данных на основе онтологий // Инженерный вестник Дона, 2017, №1. URL: ivdon.ru/ru/magazine/archive/n1y2017/3978.

2. Chittilappilly A.I., Chen L., Amer-Yahia S. A Survey of General-Purpose Crowdsourcing Techniques // IEEE Transactions on Knowledge and Data Engineering, 2016, Vol. 29, №9. pp. 2246–2266.

3. Bocharov S., Alexeeva D., Granovsky et al. Crowdsourcing morphological annotation // Computational Linguistics and Intellectual Technologies: Papers from the Annual conference “Dialogue”. Volume 1 of 2. Main Conference Program. Bekasovo: RGGU, 2013. pp. 109–114.

4. Ignatov D., Kaminskaya A., Bezzubtseva A. et al. FCA-Based Models and a Prototype Data Analysis System for Crowdsourcing Platforms // Conceptual Structures for STEM Research and Education, ICCS-ConceptStruct 2013. 20th International Conference on Conceptual Structures, ICCS 2013, Mumbay, India, January 10-12, 2013, Proceedings. Springer Berlin Heidelberg, 2013. pp. 173–192.

5. Lanser B., Unger C., Cimiano P. Crowdsourcing Ontology Lexicons // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia: European Language Resources Association (ELRA), 2016. pp. 3477–3484.

6. Киселёв Ю.А. Анализ метода выявления синонимических рядов, соответствующих одинаковым понятиям // Инженерный вестник Дона, 2015, №3. URL: ivdon.ru/ru/magazine/archive/n3y2015/3096.

7. Panchenko A., Ustalov D., Arefyev N. et al. Human and Machine Judgements for Russian Semantic Relatedness // Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Yekaterinburg,

Russia, April 7-9, 2016, Revised Selected Papers. Cham, Germany: Springer International Publishing, 2017. pp. 221–235.

8. Ustalov D. Expanding Hierarchical Contexts for Constructing a Semantic Word Network // Computational Linguistics and Intellectual Technologies: Papers from the Annual conference “Dialogue”. Volume 1 of 2. Computational Linguistics: Practical Applications. Moscow: RSUH, 2017. pp. 369–381.

9. A. Panchenko, D. Ustalov, S. Faralli et al.Improving Hypernymy Extraction with Distributional Semantic Classes // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA), 2018. URL: arxiv.org/abs/1711.02918.

10. McAndrew T.C., Guseva E.A., Bagrow J.P. Reply & Supply: Efficient crowdsourcing when workers do more than answer questions // PLOS ONE, 2017, Vol. 12, №8. pp. 1–21.

References

1. Efremova O.A., Abdullina Y.I., Yuferova A.V. Inženernyj vestnik Dona (Rus), 2017, № 1. URL: ivdon.ru/ru/magazine/archive/n1y2017/3978.
2. Chittilappilly A.I., Chen L., Amer-Yahia S. IEEE Transactions on Knowledge and Data Engineering, 2016, Vol. 29, № 9. P. 2246–2266.
3. Bocharov V., Alexeeva S., Granovsky D. et al. Computational Linguistics and Intellectual Technologies: Papers from the Annual conference "Dialogue". Volume 1 of 2. Main Conference Program. Bekasovo: RGGU, 2013. pp. 109–114.
4. Ignatov D., Kaminskaya A., Bezzubtseva A. et al. Conceptual Structures for STEM Research and Education, ICCS-ConceptStruct 2013. 20th International Conference on Conceptual Structures, ICCS 2013, Mumbay, India, January 10-12, 2013, Proceedings. Springer Berlin Heidelberg, 2013. pp. 173–192.

5. Lanser B., Unger C., Cimiano P. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia: European Language Resources Association (ELRA), 2016. pp. 3477–3484.
6. Kiselev Y.A. Inženernyj vestnik Dona (Rus), 2015, №3. URL: ivdon.ru/ru/magazine/archive/n3y2015/3096.
7. Panchenko A., Ustalov D., Arefyev N. et al. Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7-9, 2016, Revised Selected Papers. Cham, Germany: Springer International Publishing, 2017. pp. 221–235.
8. Ustalov D. Computational Linguistics and Intellectual Technologies: Papers from the Annual conference “Dialogue”. Volume 1 of 2. Computational Linguistics: Practical Applications. Moscow: RSUH, 2017. pp. 369–381.
9. Panchenko I.A., Ustalov D., Faralli S. et al. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA), 2018. URL: arxiv.org/abs/1711.02918.
10. McAndrew T.C., Guseva E.A., Bagrow J.P. PLOS ONE, 2017, Vol. 12, №8. pp. 1–21.