Внедрение в TXM дополнительных инструментов автоматической обработки текста

Alexei Lavrentiev, Fedor Solovyev, Andrey Chepovskiy

А. М. Лаврентьев
A. M. Lavrentiev
Ф. Н. Соловьев
F. N. Solovyev
А. М. Чеповский
A. M. Chepovsky

ВНЕДРЕНИЕ В TXM ДОПОЛНИТЕЛЬНЫХ ИНСТРУМЕНТОВ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТА[1]
IMPLEMENTATION IN THE TXM PLATFORM OF ADDITIONAL INSTRUMENTS OF AUTOMATIC TEXT PROCESSING

Аннотация. В докладе представлен опыт расширения возможностей платформы TXM за счет инструментов автоматической обработки текста (выделение псевдооснов, именных групп, анализ глагольного управления). В сочетании со стандартными функциями TXM (факторный анализ соответствий, специфичность и т.д.) они позволяют более эффективно осуществлять анализ специализированных корпусов, нацеленных, в частности, на выявление противоправного дискурса.

Ключевые слова. автоматический анализ текстов, платформа TXM, псевдоосновы, именные группы, глагольное управление.

Abstract. This paper presents an experience of extending the capacities of the TXM platform by adding tools of automatic text processing (allocation of pseudo-bases by stemming technique that uses a word structural pattern method, noun phrases, the analysis of verbal dependencies). Combined with the standard TXM functions (the factorial correspondence analysis, specificity, etc.) they allow the users to improve the performance of analysis of specialized corpora, such as those aimed at the detection of unlawful discourse.

Keywords. automated text analysis, TXM platform, stemming, noun phrases, verbal dependencies

Введение

В настоящей работе мы опираемся на программный комплекс – платформу TXM (http://textometrie.org). Платформа TXM является эффективным средством корпусного анализа, позволяющим проводить комплексный анализ корпусов (анализ соответствий, кластеризация, построение лексических таблиц, поиск сложных лексических конструкций, выделение подкорпусов по различным параметрам). Платформа TXM интегрирована с расширением TreeTagger [Schmid 1994], позволяющим проводить лишь морфологический анализ и лемматизацию словоупотреблений. Она использует словоупотребления в качестве структурных единиц анализа.

Для повышения эффективности таких используемых TXM методов, как анализ специфичности и анализ соответствий, целесообразно ввести в рассмотрение новые единицы анализа, опирающиеся на процедуры автоматизированной обработки текстов на естественных языках, описанные в [Чеповский 2015].

Мы предлагаем ряд расширений, позволяющих дополнить и усложнить анализ корпусов, включающий: автоматический морфологический анализ словоформ и приведение их к канонической форме, выделение псевдооснов, выделение именных и глагольных групп и комбинирование результатов работы предлагаемых расширений. Конечной целью дополнений к платформе TXM является создание механизмов для исследования применимости различных дифференцирующих признаков при решении задачи классификации текстов и создания тематических корпусов текстов.

В [Лаврентьев и др. 2018] мы провели эксперименты по использованию псевдооснов и именных групп для выявления экстремистской направленности текстов. В данной работе к этим характеристикам добавлены возможности учета глагольного управления.

Псевдоосновы

1 Работа выполнена при финансовой поддержке РФФИ в рамках научных проектов № 16-29-09546, № 18-00-00606(18-00-00233) и № 19-07-00806.

Для определения дифференцирующих признаков коротких текстов сети интернет, характеризующимися особыми тематическими и психолингвистическими свойствами, текстов, содержащих неологизмы и жаргонизмы, большой интерес представляет использование аналитического метода выделения псевдооснов, так как он позволяет обрабатывать отсутствующие в стандартных словарях формы.

Используемый способ выделения псевдооснов представляет собой метод структурных схем, описанный подробно в [Egorova и др. 2016]. Суть метода состоит в получении псевдоосновы словоформы путем рассмотрения и отбрасывания ее словоизменительных аффиксов. Словообразовательные аффиксы считаются в рамках этого метода элементом корневой части и не отбрасываются. Далее под аффиксами мы будем понимать исключительно словоизменительные аффиксы. Каждому слову можно сопоставить отвечающую ему последовательность аффиксов. Такие последовательности называются структурами некорневой части слова. Отсюда происходит название метода. Как и в традиционном морфологическом анализе, аффиксы подразделяются на префиксы и суффиксы в соответствии с их позицией относительно корня слова. Псевдоосновой называется часть слова, не содержащая суффиксов и префиксов. Способ автоматического выделения псевдооснов состоит в сопоставлении рассматриваемой словоформы с множеством допустимых в языке структур некорневой части слова. Псевдооснова слова выделяется отбрасыванием всех соответствующих определенной структурной схеме аффиксов (то есть допустимой в данном языке максимальной комбинации префиксов и суффиксов). У глаголов, в частности, отбрасываются показатели лица, числа, рода, времени, причастной формы. Видовые префиксы не отбрасываются, так как они могут влиять на лексическое значение слова.

Псевдооснова не всегда совпадает с основой слова в традиционном понимании. Например, в словоформе *людьми* единственным аффиксом, который можно отбросить согласно продуктивной структурной схеме, является *–и*, поэтому выделяется псевдооснова *людьм.*

Данный подход позволяет анализировать текстовые конструкции, опираясь не только на точные словоформы и тем самым повышает полноту и гибкость корпусного анализа.

Морфологические характеристики

Возможность привести словоформу к канонической форме позволяет анализировать различные элементы словоизменительной парадигмы как одну и ту же структурную единицу текста. Это, в свою очередь, позволяет более корректно проводить содержательный статистический анализ текста, например, путем рассмотрения частот лексем вместо частот отдельных словоформ.

При предобработке всех русскоязычных текстов мы осуществляем автоматический морфологический анализ словоформ на основе словарной компьютерной морфологии, описанной в [Чеповский 2015]. Используемая стандартная в отественной компьютерной лингвистике морфологическая модель относит каждое слово к одному из 24 морфологических классов, включающих, помимо частей речи в традиционном понимании, такие разряды, как «неизменяемое слово», «аббревиатура», «топоним». Каждый из этих морфологических классов характеризуется набором грамматических характеристик: род, падеж, число, наклонение и др. В программной реализации словарной морфологии русского языка применяется специализированная структура данных, позволяющая осуществлять поиск словоформ за линейное по числу букв словоформы время. Каждая словоформа содержит свои грамматические характеристики и её каноническую (начальную) форму.

В настоящей работе мы также использовали интегрированный в TXM программный пакет TreeTagger [Schmid 1994], предоставляющий возможность совместного морфологического анализа слов предложения на основе статистической модели, путем сопоставления словоупотреблений, снабжённых специальными метками, кодирующими морфологические характеристики. Преимуществом данной процедуры разметки является однозначность морфологического анализа, но при таком анализе существует риск ошибок, который возрастает, если текст содержит большое количество неологизмов и нестандартных написаний слов. Все виды морфологической разметки использовались в дальнейшем для сопоставительного анализа текстов корпуса.

Выделяемые из текста конструкции.

Дополнительную информацию о специфическом содержании текста можно почерпнуть, анализируя не только словоформы, но и целые именные группы. Именная группа определяется нами как группа слов, у которой главное слово существительное, а другие слова связаны с ним подчинительными синтаксическими связями. Рассмотрение частотных именных групп и их сочетаний, в совокупности с анализом отдельных словоупотреблений позволяет получить более полную картину семантических и стилистических характеристик текста, релевантных для его содержания.

Определенную сложность при выделении именных групп представляет множественность морфологических разборов при омонимии. В ходе анализа слов в предложении наш метод предполагает рассмотрение всего множества возможных морфологических разборов каждого слова.

Используемый нами алгоритм подробно описан в [Чеповский 2015] и анализирует предложения русского языка в три этапа: 1) установление подчинительных синтаксических связей в предложении между парами слов; 2) установление синтаксических связей внутри конструкций с однородными членами; 3) выделение именных групп как цепочки последовательно связанных подчинительными связями слов.

Выделение глагольных групп (словосочетаний, главным словом которых является глагол), установление связей выделенных именных групп с глаголами представляет важную, необходимую составляющую синтаксического анализа предложения. Данные задачи решаются анализом глагольного управления в рамках коммуникативной грамматики. В рамках нашей работы был использован электронный словарь глагольного управления, в который вошли первые две тысячи наиболее частотных глаголов русского языка по материалам Национального корпуса русского языка (ruscorpora.ru).

Словарь глагольного управления содержит набор ограничений, сопоставленных глаголу, и образует парадигму глагольного управления для данного глагола. Глагольным управлением является языковое явление, состоящее в проистекающих из семантики глагола требованиях, накладываемых последним на зависимые от него слова. Именно эти требования мы формализуем в виде указанных ограничений.

Результаты морфологического анализа и процедуры выделения именных групп позволяют, используя словарь глагольного управления, выявить синтаксические связи для определения глагольных групп. Выделение глагольных групп в предложении осуществляется путем анализа всех возможных пар (глагол, именная группа) предложения на предмет соответствия именной группы парадигме управления соответствующего глагола и принятия решения о наличии управления именной группы глаголом.

Анализ подкорпусов

Удобным инструментом количественной оценки «необычности» специального подкорпуса относительно всего корпуса является показатель специфичности [Lafon 1980]. Анализ специфичности позволяет составить своего рода «профиль» подкорпуса, выделенного на каких-либо внешних основаниях (например, автор, жанр, тематика или идеологическая направленность текста) путем выявления наиболее характерных или нехарактерных для него словоформ (лексем, псевдооснов, именных и глагольных групп и т.п.). Этот «профиль» может быть использован для диагностики нового текста.

Другим подходом к анализу разделенного на части (подкорпуса) по определенному критерию корпуса является анализ соответствий. Методика анализа соответствий, используемая ТХМ, была предложена Ж.-П. Бензекри [Benzecri 1979] и имплементирована в пакете FactoMineR для платформы R [Lê et al. 2008]. Анализ соответствий демонстрирует взаимную «близость» или «удаленность» подкорпусов на основе анализа частот совместного появления значений переменных (словоформ, начальных форм, псевдооснов, именных групп, морфологических тегов и т.д.).

Экспериментальный корпус был проанализирован с использованием двух обозначенных выше функций ТХМ – специфичность и анализ соответствий. Детально были рассмотрены следующие лексические объекты: словоформы; начальные формы слов, полученные по словарной морфологии;

начальные формы слов с морфологическими характеристиками, полученные с помощью TreeTagger; псевдоосновы слов; именные группы, составленные из словоформ; именные группы, составленные из начальных форм; именные группы, составленные из псевдооснов вместо отдельных словоупотреблений; глагольные группы.

Заключение

Проведенная работа по интеграции инструментов автоматической обработки текста и платформы корпусного анализа TXM показал, что такая интеграция позволяет расширить возможности статистического анализа текстов.

Детально были рассмотрены такие лексические объекты, как леммы, псевдоосновы, именные и глагольные группы различной структуры. Упомянутые средства были объединены в набор утилит, позволяющих вычислять для текстовых корпусов ряд характеристик языковых единиц, входящих в их состав. Корпуса с вычисленными характеристиками преобразуются нами в формат для импорта пакетом TXM.

Показано, что при делении текстов на подкорпуса, есть возможность интерпретировать близость, или разделенность значений рассматриваемых характеристик подкорпусов относительно друг друга как оценку, указывающую на сходство или различие маркированных подкорпусов между собой и по отношению к «нейтральному» подкорпусу.

В силу выявленных особенностей и противопоставленности нейтрального подкорпуса остальным, сформированный корпус может быть использован для машинного обучения в задачах классификации текстов на предмет выявления заданного содержания с целью их углубленного экспертного анализа.

В ходе дальнейших исследований мы планируем провести широкие исследования влияния различных дифференцирующих признаков и их комбинаций для формирования специализирующих подкорпусов текстов, наборов применяемых методов статистического и качественного анализа корпусов, формирования обучающих выборок и решения задач классификации текстовых массивов.

Литература

1 *Лаврентьев А. М., Смирнов И. В., Соловьев Ф. Н., Суворова М. И., Фокина А. И., Чеповский А. М.* (2018), Создание специальных корпусов текстов на основе расширенной платформы TXM, Системы высокой доступности, 14 (3), с. 76–81.

2. *Чеповский А. М.* (2015), Информационные модели в задачах обработки текстов на естественных языках. Второе издание, переработанное. М.

3. *Benzécri J.-P.* (1979), L'analyse des données: l'analyse des Correspondances. 2nd ed., vol. 2. Paris.

4. *Egorova E., Chepovskiy A., Lavrentiev A.* (2016), A structural pattern based method for automated morphological analysis of word forms in a natural language, Journal of Mathematical Sciences, 214 (6), pp. 802-813.

5. *Lafon P.* (1980), Sur la variabilité de la fréquence des formes dans un corpus, Mots, 1, pp. 127–165.

6. *Lê S., Josse J., & Husson F.* (2008), FactoMineR: an R package for multivariate analysis, Journal of statistical software, 25 (1), pp. 1–18.

7. *Schmid H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees, Proceedings of International Conference on New Methods in Language Processing. Manchester, UK, available at http://www.cis.uni-muenchen.de/sschmid/tools/TreeTagger/data/tree-tagger1.pdf

References

1. *Benzécri J.-P.* (1979), L'analyse des données: l'analyse des Correspondances. [Data Analysis: Correspondence Analysis] 2nd ed., vol. 2. Paris.

2. *Chepovskiy A. M.* (2015), Informatsionnye modeli v zadachakh obrabotki tekstov na estestvennykh yazykakh [Information models in natural language text processing problems], 2nd ed. Moscow.

3. *Egorova E., Chepovskiy A., Lavrentiev A.* (2016), A structural pattern based method for automated morphological analysis of word forms in a natural language, Journal of Mathematical Sciences, 214 (6), pp. 802-813.

5. *Lavrentiev A. M., Smirnov I. V., Solov'ev F. N., Suvorova M. I., Fokina A. I., Chepovskiy A. M.* (2018), Sozdanie specialnikh corpusov tekstov na osnove rasshirennoy platformi TXM [Creating text corpora for special purposes on the basis of extended TXM platform], Systemi Visokoy Dostupnosti [Highly Available Systems], 14 (3), pp. 76–81.

6. *Lafon P.* (1980), Sur la variabilité de la fréquence des formes dans un corpus [On the Variability of Word-Form Frequencies in a Corpus], Mots, 1, pp. 127–165.

7. *Lê S., Josse J., & Husson F.* (2008), FactoMineR: an R package for multivariate analysis, Journal of statistical software, 25 (1), pp. 1–18.

8. *Schmid H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees, Proceedings of International Conference on New Methods in Language Processing. Manchester, UK, available at http://www.cis.uni-muenchen.de/sschmid/tools/TreeTagger/data/tree-tagger1.pdf