УДК 004.91

ВИЗУАЛИЗАЦИЯ ПОСЕССИВНЫХ КОНСТРУКЦИЙ С ТЕРМИНАМИ РОДСТВА В ВИДЕ ГЕНЕАГОГИЧЕСКИХ ДЕРЕВЬЕВ

Голуб А.Л., Белова А.С., Бондаренко Г.В.
Научный руководитель – к. т. н., ординарный доцент ф-та ИКТ Университета ИТМО Коцюба И.Ю.
Федеральное государственное автономное образовательное учреждение высшего образования “Национальный исследовательский университет ИТМО”

Аннотация
Посессивные конструкции с терминами родства в русском языке (например, *муж тёщи моей сестры*), будучи частотны в текстах различных жанров, в подавляющем большинстве случаев становятся проблемой для человеческого восприятия. В статье описаны принципы работы созданной авторами компьютерной программы, способной выделять такие конструкции в тексте и визуализировать их в удобном для восприятия виде. Такая попытка автоматизации анализа родственных связей в тексте способна внести вклад в развитие исторической науки, лингвистики и литературоведения.

Введение

Посессивные конструкции с терминами родства (например, *муж тёщи моей сестры*, *бабушка деверя снохи его брата*), будучи частотны в текстах различных жанров, в подавляющем большинстве случаев становятся проблемой для человеческого восприятия. Целью данного проекта было создание программы, способной выделять такие конструкции в тексте и визуализировать их в виде генеалогических деревьев. Такая попытка автоматизации анализа родственных связей в тексте способна внести вклад в развитие исторической науки, лингвистики и литературоведения, в частности в изучение текстов художественной литературы и личных дневников.

Среди рассмотренных решений аналогичных задач отсутствует готовый инструмент, позволяющий в полной мере достичь поставленной цели. В отдельности грамматика посессивных конструкций и системы терминов родства в различных языках хорошо изучены в лингвистике, однако посессивные грамматические конструкции, содержащие термины родства, не имеют достаточной глубины рассмотрения. В свою очередь, для изображения генеалогических деревьев существует множество инструментов визуализации. Используются библиотеки для построения графов (например, Graphviz), а также специализированные программные пакеты (ggenealogy). Однако данные инструменты не соответствуют в полной мере специфике поставленной задачи: иерархическая структура данных не допускает наличия горизонтальных связей, что требуется для корректного отображения отношений. Также существует программное обеспечение, позволяющее составлять генеалогические деревья в интерактивном режиме (Family Historian). Достоинством таких сервисов является наглядность, но из-за обязательного взаимодействия с пользователем построение графов проблематично автоматизировать. В связи с этим мы разработали собственный алгоритм визуализации, удовлетворяющий всем установленным требованиям.

Основная часть

Код проекта написан на языке Python. На вход дается текст, в котором ищутся посессивные грамматические конструкции, состоящие из терминов родства, и, возможно, других существительных, притяжательных прилагательных и притяжательных местоимений. Поиск осуществляется при помощи Python-библиотеки для обработки естественного языка NLTK и морфологического анализатора pymorphy2. Далее каждая из найденных конструкций подготавливается для анализа указанных в ней родственных связей. Для этого термины родства приводятся к начальной форме, притяжательные прилагательные заменяются на

форму единственного числа именительного падежа существительных, от которых они образованы, а притяжательные местоимения – на соответствующие им личные также в именительном падеже. Существительные, не являющиеся терминами родства, приводятся к форме именительного падежа с сохранением числа. Затем слова выписываются в порядке родства; см. пример: *Того ученика я взял не просто с улицы — это был* ***племянник моей жены***.
=> ***я жена племянник***

Далее для каждого термина родства в последовательности из заранее созданного файла загружается шаблон фрагмента генеалогического древа, отображающий родственную связь между персонажами, соответствующими данному и предыдущему слову в конструкции. В рамках шаблона персонажи связаны друг с другом напрямую одним из двух возможных типов связи: родитель-ребенок или супруг-супруга. Затем шаблоны последовательно соединяются друг с другом в единый граф, где все персонажи также оказываются связаны напрямую одним из указанных типов связи. Следует отметить, что одному термину родства может соответствовать более одного шаблона вследствие языковой неоднозначности; в таком случае для данной посессивной конструкции будет создано несколько вариантов графа.

Для визуализации графа используются Python-библиотеки NetworkX и Matplotlib. Форма узлов соотнесена с полом персонажей: круг для женского, квадрат для мужского; для первого персонажа в конструкции используется ромб. Узлы первого и последнего персонажа в конструкции окрашиваются в синий цвет, а узлы остальных персонажей, упомянутых в ней напрямую, - в голубой. Персонажи, на которых нет прямого указания, изображаются серым цветом. Для каждого узла задаются координаты: родители рисуются на один пункт выше детей, супруги - на одном уровне; если полученная координата уже занята, узел изображается на один пункт правее. В результате создается файл формата PNG с изображением графа, а также указанием анализируемой конструкции и предложения в тексте, откуда она взята.

Выводы

На данный момент программа способна находить в тексте посессивные конструкции и визуализировать их в виде генеалогических деревьев. При тестировании на специально составленном корпусе текстов поиск конструкций осуществлялся с точностью (precision) 98% и полнотой (recall) 93%, а анализ и отрисовка конструкций - с точностью (accuracy) 95%. В дальнейшем планируется улучшить качество поиска конструкций в тексте и усовершенствовать методы анализа конструкций на предмет родственных связей. Кроме того, возможна разработка мобильного приложения с аналогичной функциональностью для широкого пользования.

Голуб А. Л. (автор) Подпись

Коцюба И. Ю. (научный руководитель) Подпись