Программа выделения русских индивидуализированных именных групп **TagLite**

TagLite: The Program of Identification of Russian Individualized NPs

Л.Г. Крейдлин

ИППИ РАН, Москва

lenya@iitp.ru

В статье описывается алгоритм программы, предназначенный для автоматического распознавания и выделения в текстах на русском языке именных групп с именами собственными, которые обозначают некоторые индивидуализированные объекты, а именно людей, учреждения и географические объекты.

1. Введение

Автоматическое распознавание и последующее выделение в текстах именных групп с собственными именами, обозначающих людей, учреждения и географические объекты, является особой проблемой компьютерной обработки текстов на естественном языке (АОТ). Имена собственные, в отличие от имен нарицательных, образуют открытый, постоянно расширяющийся список. Поэтому пополнение ими словаря системы, предназначенной для автоматической обработки текстов (АОТ), в значительной степени осложнено.

Мы исходим из того, что АОТ-система должна уметь обращаться с каждой новой единицей как с уже известными, то есть быть способной отнести конкретную единицу к одному из классов, заданных в системе, и приписать ей необходимые признаки. Одновременно необходимо, чтобы система АОТ, в особенности та, в задачу которой входит обработка больших текстовых массивов, могла самостоятельно интерпретировать отдельные неизвестные слова и синтаксически связанные группы слов и классифицировать их. Такая проблема особенно остро встает при обработке текстов, изобилующих названиями организаций или географических объектов.

2. Несколько слов об идентификации текстовых объектов

В настоящей работе мы ставили перед собой задачу создать действующий алгоритм и программу распознавания в текстах именных групп, которые обозначают индивидуализированные объекты трех типов:

1) “человек” (*финансовый директор ЮКОСа Брюс Мизамор*);

2) “географический объект” (*город Ногинск Костромской области*);

3) “организация” (*стадион им. Кирова*).

Решение этой задачи связано с проблемой идентификации имен собственных и просто цепочек обычных слов, которые ведут себя в текстах как имена собственные. Таким образом, выделению подлежат как имена *Айон Неприген* или *“Импэксбанк”*, так и названия *“Аэрофлот – российские авиалинии”, “Ярославский шинный завод”*. Помимо распознавания таких ИГ, АОТ-система должна осуществлять их семантическую классификацию.

Обнаружение в тексте индивидуализированных именных групп (ИИГ) является более широкой задачей, чем выделение и интерпретация отдельных имен и названий сущностей (англ. *named entities*, далее NE). Причина этого в том, что ИИГ могут состоять не только из NE: имена людей, названия учреждений и географических объектов – это обязательный, но не единственно возможный компонент в составе ИИГ.

Насколько нам известно, сегодня не существует масштабных АОТ-систем, способных вычленять и маркировать ИИГ в текстах на русском языке. Тем не менее, описанию разнообразных систем идентификации NE посвящено довольно большое число зарубежных публикаций. Их авторы предлагают различные методы распознавания и смысловой интерпретации, которые можно условно разделить на следующие группы:

- статистический подход (для создания статистической модели используется корпус размеченных текстов, см. Sekine S., Eriguchi Y. [2000]);
- вычислительные методы на основе обучающих моделей (например, в рамках проекта CoNLL-2003);
- метод контекстного анализа (опирающийся на правила идентификации NE в тексте в зависимости от левого и правого контекста и

списки слов открытых классов, см. McDonald D. [1996], Kokkinakis D. [2004]);

- гибридный подход (объединяющий статистические методы и приемы контекстного анализа (см. Mikheev A. et al. [1998]).

Из-за отсутствия большого семантически аннотированного корпуса русских текстов, на основе которого можно было бы создать надежную статистическую модель обработки текстов и автоматического выделения ИИГ, мы для решения поставленной задачи избрали метод контекстного анализа.

Среди близких нам работ отметим, прежде всего, статью McDonald D. [1996]. В ней описывается один из ключевых компонентов системы понимания естественного языка Sparser – модуль PNF, который предназначен не только для распознавания и классификации собственных имен (имена и названия, несмотря на все их многообразие, имеют, по словам автора, более или менее регулярную внутреннюю структуру, которую можно выделять, используя правила контекстно-зависимой грамматики), но и используется для вычленения и интерпретации именных групп.

В статье выделяются три этапа работы, так или иначе связанные с выделением и обработкой собственных имен: 1) определение границ последовательности слов, из которых образуется имя собственное; 2) отнесение полученного элемента к той или иной семантической категории с одновременным разрешением неоднозначности; и 3) сохранение полученного результата в модели с целью его дальнейшего использования при работе Sparser как с данным и другими текстами.

Лексико-грамматические сведения о словах и цепочках слов становятся базой, позволяющей выявить структуру в пределах ИИГ, и являются основой для последующего отнесения ИГ к той или иной категории. Ключевыми для классификации собственных имен в PNF являются топонимы (ср.: пример автора ***Cambridge*** *Savings Bank*); слова, обозначающие социальные, религиозные и иные учреждения (*church* или *bank*); аббревиатуры, используемые в обозначении компаний (*Inc., Ltd.*), титулы (*Dr.*, *Mr.*) и др. Показателями индивидуализированных именных групп служат встречающиеся в тексте и уже известные системе имена собственные, а также слова-классификаторы (*spokesman, company*), которые обеспечивают возможность прогнозирования: непосредственно после таких лексических единиц велика вероятность встретить в тексте имя собственное.

Отправной точкой работы Kokkinakis D. [2004] служит тезис о том, что (газетные) тексты изобилуют названиями объектов и именами (часто представленными в виде последовательностей из нескольких слов), к которым не применимы «традиционные правила» выделения NE. Автор указывает, что именные сущности могут иметь сложную внутреннюю структуру: так, собственное имя может состоять исключительно из обычных слов (*The World Intellectual Property Organization*). Он также справедливо замечает, что NE совершенно необязательно должны иметь лексическое значение (ср.: *F117A* [модель самолета], SN 1987 A [сверхновая звезда]).

В работе представлена шведская система распознавания NE (NER), использующая более детальную, нежели у других авторов, классификацию NE (8 основных категорий и 47 подтипов). NER состоит из следующих основных компонентов: 1) списки имен, состоящих из одного (всего около 95 тыс. входов) или нескольких слов; 2) система поверхностного грамматического разбора – конечный автомат, построенный на базе правил контекстно-зависимой грамматики; 3) модуль анализа информации о текстовых сущностях, полученной от двух предыдущих компонентов, и классификации тех единиц, которые не были проинтерпретированы; 4) система маркировки NE в выходном тексте, которая осуществляет проверку разметки и устраняет возможные ошибки в ней (если NER не удалось до сих пор определить категорию какого-либо текстового элемента, этот модуль может ее угадать и правильно пометить, запустив процедуру проверки орфографии или анализа более широкого контекстов ранее классифицированных NE).

В статье Stevenson M., Gaizaukas R. [2000] обсуждается серия экспериментов с системой, которая была построена авторами на основе комплекса LASIE (разные его версии использовались в проектах MUC и HUB4). Целью этих экспериментов была идентификация в текстах собственных имен на основе предварительно построенных списков слов и словосочетаний. Описываемая авторами система представляет интерес, поскольку она является самообучающейся: пользуясь достаточно простым набором фильтров, программа пополняет ранее построенные списки имен новыми единицами, в результате чего она становится гораздо более точной. В работе излагаются способы построения списков, их пополнения, а также описываются фильтры и методы экспериментальной работы с ними.

3. `TagLite`: структура и алгоритм работы

Представляемая ниже программа `TagLite`, или **тэггер**, ищет во входном тексте индивидуализированные именные группы и в соответствии с результатами поиска осуществляет его разметку. Обработка текста при помощи `TagLite` происходит в несколько. Основными из них являются: членение текста на отдельные (значимые) единицы, морфологический анализ текста, обращение к спискам слов, контекстный анализ, разрешение неоднозначности и собственно разметка.

Алгоритм, лежащий в основе программы, использует линейное, а не структурное

представление текста. При разбиении текста TagLite собирает разнообразную информацию о каждом слове: для программы существенно, написано ли оно с большой буквы, заключено ли в кавычки и какие знаки препинания его окружают. Далее текст подается на вход морфологического анализатора, который строит его линейное морфологическое представление, соотнося каждое вхождение слова с его исходной формой и приписывая ему набор (наборы) релевантных морфологических характеристик[1].

Построив морфологическое представление, TagLite обращается к заранее построенным спискам слов, которые либо могут входить в состав ИИГ, либо являются собственными именами. Каждая единица входного текста проверяется на предмет вхождения в списки, для составления которых нам пришлось решать задачу построения не вполне традиционной классификации собственных имен.

Выделяются три основные группы слов, помещаемых в списки. Это **имена собственные**, **ИГ-образующие** слова и **ИГ-зависимые**. Имена собственные делятся на антропонимы и топонимы. Из антропонимов учитываются только личные имена, фамилии и отчества людей.

Вторая группа слов – классификаторы индивидуализированных именных групп, или, иначе, ИГ-образующие. Это имена нарицательные, которым свойственно употребление в одном контексте с именами собственными и которые в норме играют роль синтаксической вершины ИИГ, как, например, выделенные слова в сочетаниях ***монтер*** *Семен Петров* и ***озеро*** *Чертог*. ИГ-образующие мы разделяем на несколько классов в соответствии с их семантикой. Это (1) слова, обозначающие различные характеристики людей – этнические (*черкешенка*), социальные (*посол*), оценочные (*зануда*); (2) различные социальные организации и учреждения (*институт, контора*) и (3) имена, которые вводят в текст топонимы и обозначают их местоположение или географические ориентиры (*река, окраина*). В первый класс входит огромное и разнородное множество лексических единиц, а потому для начала было решено ограничиться названиями профессий.

Мы создали несколько списков с ИГ-зависимыми словами, то есть именами, входящими в ИИГ, но не являющимися их вершиной. Речь идет о порядковых прилагательных (***первый*** *президент США*), прилагательных, обозначающих место определенного лица в некой последовательности или должностной иерархии (***бывший*** *премьер-министр Михаил Касьянов,* ***старший*** *оперуполномоченный Цветков*), прилагательных, семантически связанных с названиями топонимов (***архангельский*** *священник Владимир Купленский*) или являющихся топонимами (***Соловецкий*** *монастырь*). Кроме того, нами были составлены списки аббревиатур, инициалов, показателей титулов (ср.: ***фон*** *Караян*), типовых основ (*авиа, пресс)*, при помощи которых образуются сложные слова (композиты), и др.

Исследователи, занимающиеся компьютерной обработкой текстов, в большинстве своем сходятся во мнении, что использование в АОТ-системах различных баз знаний, энциклопедических данных, тезаурусов и другой вспомогательной информации приводит к значительно лучшим результатам при решении задач по автоматической синтаксической и семантической разметке текста, поиску, извлечению из текста информации и разрешению омонимии. У нас такой вспомогательной информацией являются списки слов[2].

Если слово обнаружено в одном из списков, то информация о принадлежности к списку приписывается слову как отдельный признак, называемый **статусом** слова. Если слово присутствует сразу в нескольких списках, оно приобретает несколько статусов. Так, *Владимир* – это и название города, и имя человека; соответственно, данная единица получает две пометы – **FNAME** (“имя”) и **GEOGR** (“топоним”).

Для дальнейшей работы программа TagLite должна проинтерпретировать слова, которые прежде не были идентифицированы ни с одной из записей в списках, а также единицы, не опознанные морфологическим анализатором (как правило, эти множества пересекаются). Определение статуса слова в таких случаях достигается на следующем этапе при работе **гессера**.

Гессер состоит из нескольких частей, позволяющих проанализировать слово как (а) аббревиатуру или композит; (б) слово, имеющее типовое именное окончание. Помимо этого, гессер позволяет (в) установить статус ранее неопознанного слова в зависимости от контекста и, если слову было приписано несколько статусов, снять неоднозначность, опять-таки опираясь на контекстный анализ. Для этого в тэггере предусмотрены специальные правила, областью действия которых являются не лексические

[1] Система морфологического анализа, к которой обращается программа TagLite, была разработана в Лаборатории компьютерной лингвистики ИППИ РАН Н.В. Григорьевым. Используемый данной системой словарь мы скомпилировали из морфологического словаря лингвистического процессора ЭТАП-3.

[2] TagLite обращается в общей сложности к 16 спискам, для компиляции которых мы использовали данные комбинаторного словаря системы ЭТАП-3, а также ряд обычных, таких как Шведова Н.Ю. [1998-2000], и онлайновых лексикографических источников. В результате, только в списке фамилий насчитывается около 19 тысяч записей. Все списки остаются открытыми и дополняются без изменений структуры и кода тэггера.

элементы входного текста, а их статусы. Эти правила, а их в программе более 100, преобразуют одну последовательность статусов в другую при заданных контекстных условиях.

Устранение неоднозначности статуса у некоторой единицы нередко ведет к определению статуса разных форм одного и того же слова. Пусть на вход программе поступил текст *Николаев* (GEOGR LNAME) *спал. Если бы не тяжелая работа, Петру* (FNAME) *Николаеву* (GEOGR LNAME) *не надо было бы вставать назавтра так рано.* Здесь нас интересуют вхождения словоформ *Николаев* и *Николаеву*. Руководствуясь существующим правилом, гессер однозначно определит, что словоформе *Николаеву* следует приписать статус "фамилия" (а не "топоним"). Действуя рекурсивно, гессер вернется к вхождению *Николаев* и установит, что это фамилия человека. Опирается он при этом на принцип согласования статусов в рамках одного текста (разумеется, работая с текстом, где говорится, что некий Николаев приехал в Николаев, программа может допустить ошибку).

После того, как словам приписаны все возможные и необходимые статусы и гессер устранил все те неоднозначности, которые ему подвластны, наступает финальная часть работы программы. Она состоит из двух этапов. Сначала применяются правила, определяющие, какие последовательности статусов какими тэгами (парой специальных маркеров, показывающих начало и конец выделяемой именной группы) должны быть помечены. С этой целью маркируются начальные границы слов текста, а сами слова временно заменяются их статусами с сохранением разметки текста – отступов, абзацев, знаков препинания и т.д. Затем запись текста в виде цепочки статусов переводится обратно в запись на естественном языке. TagLite восстанавливает исходную разметку и пунктуацию текста и проставляет тэги, которые были получены на предыдущем этапе.

4. Оценка работы программы **TagLite**

Для оценки качества работы программы TagLite, нами были отобраны свободно доступные в интернете новости и статьи на общественно-политические темы. Эти материалы составили тестовый текст (5748 слов), в котором мы выделили 369 индивидуализированных именных групп. Из их числа к категории "человек" относилось 155 именных групп, а к "географическим объектам" и "организациям" – 141 и 73 ИГ, соответственно. Тем самым, мы получили эталон разметки, к которому должен стремиться тэггер.

Существует несколько общепринятых параметров, позволяющих судить о качестве работы АОТ-системы. Чаще других используются такие критерии оценки, как точность (*precision*) и полнота (*recall*). Эти критерии применялись и нами.

Пусть $\mathbf{R_e}$ есть число индивидуализированных именных групп в эталонном тексте, $\mathbf{R_t}$ – общее число ИГ, помеченных программой TagLite, а $\mathbf{N_t}$ и $\mathbf{N_g}$ – количество таких размеченных TagLite именных групп, границы которых совпадают с границами ИИГ из эталона и которым либо приписаны правильные тэги (то есть ИГ отнесена к правильной семантической категории, $\mathbf{N_t}$), либо относительно которых была выдвинута верная гипотеза о типе тэга ($\mathbf{N_g}$). Если использовать введенные обозначения, то точность разметки индивидуализированных именных групп подсчитывается по формуле $precision = \frac{N_t + N_g}{R_t}$, а ее полнота – $recall = \frac{N_t + N_g}{R_e}$. Чем ближе значения *precision* и *recall* к 1, тем лучше результаты работы тэггера. Соответственно, реальным показателем качества выделения ИИГ будет комбинация этих двух критериев оценки, которую принято называть *f-value*, и которая может быть выражена формулой:

$$f\text{-}value = \frac{2 \times precision \times recall}{precision + recall}$$

Ниже приведены значения описанных параметров одновременно для всех трех категорий индивидуализированных именных групп, которые умеет маркировать TagLite.

Все категории		
precision	***recall***	***f-value***
0,843	0,873	0,858

Аналогичным образом, нами были подсчитаны значения *precision*, *recall* и *f-value* для каждой из трех выделяемых категорий ИИГ:

	precision	***recall***	***f-value***
PERS	0,861	0,916	0,888
GEOGR	0,951	0,957	0,954
INST	0,776	0,616	0,687

Как видно из последней таблицы, хуже всего TagLite обрабатывает именные группы, обозначающие организации. Это объясняется несколькими причинами. Во-первых, в работе мы больше обращали внимание на маркирование ИИГ, отображающих людей, а потому правил для обработки ИГ из категории "организация или учреждения", в TagLite численно меньше всего. Во-вторых, индивидуализированные именные группы – названия учреждений – наиболее сложны для описания в рамках правил анализа линейного контекста. Так, если программа встречает ИГ

Дербентский завод игристых вин, то неясно, по каким основаниям тэггер может маркировать всю группу, а не одну только первую ее часть *Дербентский завод* (ср. с фразой *Дербентский завод игристых вин не производит*).

Что касается прочих результатов разметки тестового текста, отметим, что с выбором тэга (но не границ ИГ!) при обработке тестового текста программа ошиблась всего один раз, когда в тексте встретилось название организации *Аль-Гамаа аль-Исламийя*, которое было проинтерпретировано TagLite как имя и фамилия человека. Ошибки в выборе либо левой, либо правой границы именной группы носят более частотный характер – 4,07% (2,44% и 1,63%, соответственно), а доля ИИГ, границ которых тэггер распознать не сумел, составила 4,34%.

Результаты работы многих систем распознавания NE достаточно близки тем, что были получены нами. В частности, значения *precision*, *recall* и *f-value* (для трех выделяемых категорий – названий организаций, географических объектов и имен людей) в работе Stevenson M., Gaizaukas R. [2000] равнялись 0,93, 0,81 и 0,87, соответственно. Однако прямое сравнение результатов здесь не совсем корректно, поскольку перед нами стояла более широкая задача – выделение индивидуализированных именных групп, а не NE[3].

5. Заключение

В заключение укажем на “слабые” места программы TagLite и наметим возможные перспективы ее усовершенствования. Пока еще относительно невысокой является скорость обработки тэггером больших, которую, как нам представляется, можно значительно повысить за счет содержательных изменений в алгоритме TagLite и, возможно, переписав исходный код с языка Perl на C++.

Качественного улучшения следует ожидать и от построения более полной классификации анализируемых единиц, составления новых списков и расширения уже существующих. Последнюю операцию не обязательно делать вручную: в TagLite можно ввести процедуру автоматического добавления в списки ранее неизвестных слов – в случае однозначного определения их статусов. Положительных изменений в работе тэггера, в частности, гессера, можно добиться при введении правил, опирающихся на анализ глаголов, обнаруженных в контексте того или иного слова с неопределенным или неоднозначным статусом. Например, если в обрабатываемом тексте встречается фраза, начинающаяся словами *В своем интервью Тавола отметил, что…,* а из остальных частей текста не удается установить характер единицы *Тавола*, то присутствие ее в контексте глагола говорения *отметил* дает основания для изменения статуса слова *Тавола* на “имя” или “фамилия”.

[3] В принципе, в TagLite предусмотрена возможность маркирования NE, а не ИИГ, однако оценку результатов этой операции мы не проводили.

Литература

Kokkinakis D. Reducing the Effect of Name Explosion // Proceedings of 4th International Conference on Language Resources and Evaluation, 2004. С. 1-6.

McDonald D. Internal and External Evidence in the Identification and Semantic Categorization of Proper Names // Corpus Processing for Lexical Acquisition. The MIT Press, 1996. Т. 2, С. 32-43.

Mikheev A., Grover C. and Moens M. Descripton of the LTG system used for MUC-7 // Proceedings of 7th Message Understanding Conference, 1998.

Mikheev A., Grover C. and Moens M. Named Entity Recognition without Gazetteers // Proceedings of the EACL-1999, С. 1-8.

Osenova P., Kolkovska S. Combining the Named-Entity Recognition Task and NP Chunking Strategy for Robust Pre-processing // Proceedings of 1st Workshop on Treebanks and Linguistic Theories, 2002. С. 167-182

Sekine S., Eriguchi Y. Japanese Named Entity Extraction Evaluation. // Proceedings of the COLING-2000. С. 1106-1110.

Stevenson M., Gaizaukas R. Using Corpus-derived Name Lists for Named Entity Recognition // Proceedings of ANLP-NAACL-2000. С. 290-295.

Yangarber R., Grishman R. Machine learning of extraction patterns from un-annotated corpora // Proceeding of the Workshop on Machine Learning for Information Extraction, ECAI-2000.

Алексеев Д.И., Гозман И.Г., Сахаров Г.В. Словарь сокращений русского языка. 4-е изд. Под ред. Д.И. Алексеева. М.: "Русский язык", 1984.

Русский семантический словарь. Под ред. Н.Ю. Шведовой. М.: "Азбуковник", 1998-2000. Т. 1–2.