УДК 81 + 811.161.1 + 81’33
DOI 10.25205/1818-7935-2019-17-1-65-77

Использование мер семантической близости для распознавания кореференции в русском языке

И. Л. Азеркович

Национальный исследовательский университет «Высшая школа экономики»
Москва, Россия

Аннотация

Статья посвящена описанию серии экспериментов по исследованию роли семантической информации в разрешении кореферентных связей для русского языка, ее использованию в системах автоматического анализа текстов и оценке результатов их работы. Целью первого этапа экспериментов было определить, какие метрики семантической близости между референциальными выражениями больше соответствуют кореферентным связям между ними. Подсчет метрик производился на материале русской Википедии и тезауруса RuThes. На втором этапе была разработана система автоматического распознавания кореферентности, использующая метрики семантической близости в качестве признаков для машинного обучения, и оценено качество ее работы. Результаты проведенных экспериментов позволяют установить метрики семантической близости, подходящие для использования в системах разрешения кореферентности, а также демонстрируют повышение качества работы подобных систем при использовании семантической информации.

Ключевые слова

автоматическая обработка естественного языка, распознавание кореферентности, метрики семантической близости, машинное обучение, русский язык

Для цитирования

Азеркович И. Л. Использование мер семантической близости для распознавания кореференции в русском языке // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2019. Т. 17, № 1. С. 65–77. DOI 10.25205/1818-7935-2019-17-1-65-77

Using Semantic Relatedness Measures in Coreference Resolution for Russian

Ilya L. Azerkovich

National Research University Higher School of Economics
Moscow, Russian Federation

Abstract

The paper is devoted to the role of semantic information (in the form of semantic relatedness measures) in coreference resolution for the Russian language. It describes a series of experiments in calculating metrics of semantic relatedness based on Russian material and evaluating the possibility of using them in systems of natural language processing, as well as the performance of such systems.

The goal of the first stage of experiments was to find out, which semantic relatedness measures better correspond to coreference relations between referential expressions. For this purpose, several metrics calculated from different parameters were chosen and evaluated on the test set, derived from the Russian coreference corpus RuCor. Semantic data for the metrics was obtained from two sources: Russian Wikipedia and RuThes thesaurus. The results showed that while RuThes provided more reliable data for common nouns, Wikipedia data correlated better with named entities. Based on the obtained results, metrics that corresponded to coreference relations the most were chosen to be implemented during the next stage of experiments.

For the second stage of experiments a machine-learning based coreference resolution system that could use semantic relatedness measures as features was developed, based on the decision trees classification algorithm. Four versions of the system were tested: without any features derived from semantic information, with features derived from only one

of the sources, and with features derived from both sources. Tests were performed on the subset of RuCor corpus that already included gold standard mark-up as the base for evaluation. The tests showed noticeable improvement for the version that was using semantic information from both data sources.

The experiments made demonstrate the increase of quality of coreference resolution with the implementation of features based on semantic information. The results obtained are comparable to or exceed the ones described in similar papers on the topic of Russian coreference resolution.

Keywords

natural language processing, coreference resolution, semantic relatedness measures, machine learning, Russian language

For citation

Azerkovich, Ilya L. Using Semantic Relatedness Measures in Coreference Resolution for Russian. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2019, vol. 17, no. 1, p. 65–77. (in Russ.) DOI 10.25205/1818-7935-2019-17-1-65-77

Введение

Распознавание кореферентных связей является одной из задач, решение которой актуально для различных систем автоматического анализа текста, таких как, например, извлечение информации из текста, машинного перевода и др. Два языковых выражения находятся в кореферентных отношениях, если они обозначают одну и ту же сущность реального мира. Таким образом, задача распознавания кореферентных отношений состоит в выделении в тексте цепочек упоминаний одной и той же сущности. Кореферентную цепочку, например, составляют выделенные референциальные выражения в следующем примере:

(1) [**Facebook**] уже несколько лет делает ставку на видео. <…> К февралю 2017 года наконец стало понятно, насколько амбициозны планы [**компании**], – [**она**] хочет заменить собой телевизор.

Классические алгоритмы для решения такой задачи в основном опираются на данные синтаксического и морфологического разбора текста с использованием базовых семантических признаков, например информации об именованных сущностях или о семантической совместимости кореферентных выражений в терминах таксономической иерархии (ср. семантические классы, описанные в [Kameyama, 1997] или [Soon et al., 2001]). В качестве источников семантической информации во многих исследованиях используются внешние источники, такие как тезаурусы (WordNet) или онлайн-энциклопедии (Википедия) [Harabagiu et al., 2001; Ponzetto, Strube, 2006; Rahman, Ng, 2011].

Разработка систем автоматического разрешения кореферентности для русского языка началась позже, чем для английского, но ведется достаточно активно (см. описания систем в [Толпегин, 2008; Kamenskaya et al., 2014; Kutuzov, Ionov, 2014], а также [Toldova et al., 2014] о результатах соревнования между системами). Из-за морфологических и синтаксических особенностей русского языка стандартные методы, применяемые для английского языка, могут работать не так хорошо ([Toldova et al., 2016]), но, тем не менее, как показали работы [Toldova, Ionov, 2017; Azerkovich, 2018], использование признаков, основанных на семантической информации, позволяет повысить их эффективность.

Данная работа развивает результаты, достигнутые в статьях [Toldova, Ionov, 2017; Azerkovich, 2018]. В ней излагаются результаты двух серий экспериментов, в которых исследуется роль семантической информации в автоматическом распознавании кореферентности. Первая из них была призвана ответить на вопрос, какие метрики семантической близости между референциальными выражениями больше соответствуют кореферентным связям между ними. Вторая серия экспериментов описывает работу системы автоматического распознавания кореферентности, использующей выбранные на первом этапе метрики в качестве признаков для машинного обучения.

Результаты проведенных экспериментов демонстрируют повышение качества распознавания кореферентности с использованием семантической информации. Полученное улучшение превышает описанное С. Толдовой и М. Ионовым в статье [Toldova, Ionov, 2017], что

говорит о лучшем качестве использованных признаков. Кроме того, в рамках исследования проведена оценка различных мер семантической близости между референциальными выражениями и создан рабочий прототип системы автоматического распознавания кореферентности, использующий данные метрики.

Предыдущие работы

Ранние системы автоматического разрешения кореферентности были основаны на правилах, но с начала 2000-х гг. наблюдается активное развитие систем, полностью основанных на машинном обучении либо использующих комбинированный статистически-правиловый подход [Soon et al., 2001; Ng, Cardie, 2002]. Как показали результаты конференции CONLL-2012 (см. [Pradhan et al., 2012]), комбинированные системы позволяют получить лучшие результаты в сравнении с системами с другими архитектурами. Система, описанная в работе [Soon et al., 2001], считается классической благодаря простоте реализации и высокому качеству анализа и часто воспроизводится в статьях других исследователей.

Многие системы разрешения кореферентности, как ранние, так и более поздние, используют семантическую информацию в основном в виде сведений о семантической однородности (semantic consistency) референциальных выражений (например, совпадении по роду и числу или принадлежности к одному и тому же таксономическому классу, ср. [Kameyama, 1997; Harabagiu et al., 2001]). Эта информация может быть выделена как другими модулями обработки текста (например, система извлечения именованных сущностей Стэнфордского университета в [Rahman, Ng, 2011]), так и модулями, непосредственно встроенными в систему (информация о семантических ролях в [Kamenskaya et al., 2014] или извлечение именованных сущностей в [Soon et al., 2001]). С появлением электронных ресурсов, содержащих экспертную и онтологическую информацию, все большее число исследователей использует в качестве источников именно их. Одними из самых часто используемых ресурсов являются тезаурус WordNet (см. [Harabagiu et al., 2001; Ponzetto, Strube, 2006]) и онлайн-энциклопедия Википедия (см. [Ponzetto, Strube, 2006; Haghighi, Klein, 2009; Rahman, Ng, 2011]).

На материале русского языка исследований по автоматическому разрешению кореферентности и по использованию для этого семантической информации на момент написания статьи существует меньше, чем для английского. В рамках форума RuEval, прошедшего в 2014 г., было проведено первое соревнование систем такого рода [Toldova et al., 2014]. В соревновании были отдельно представлены задачи автоматического разрешения анафоры и кореферентности между именами собственными, именными группами и местоимениями, и из восьми систем – участников трека по разрешению анафоры три также приняли участие в треке по разрешению кореферентности. Однако из трех систем лишь одна [Bogdanov et al., 2014] использовала семантическую информацию, причем в качестве ее источника выступала проприетарная онтология.

Предоставление публичного доступа к корпусу текстов с кореферентной разметкой, использовавшемуся при проведении соревнования, значительно облегчило задачу разработки подобных систем, однако проблема использования семантической информации до недавнего времени была изучена недостаточно. В то же время за последний год были опубликованы две работы, описывающие улучшение качества работы систем автоматического разрешения кореферентности при использовании семантической информации: [Toldova, Ionov, 2017; Azerkovich, 2018].

Первая статья описывает эксперимент по улучшению качества распознавания кореферентности с использованием таких семантических средств, как векторы представления слов, списки именованных сущностей и тезаурус RuThes. Несмотря на сравнительно простые признаки, использовавшиеся в работе, результаты исследования демонстрировали повышение качества в пределах 0,2–0,3 %. Это позволило предположить, что использование других способов представления семантической информации позволит еще больше повысить качество распознавания. В настоящем исследовании были проведены эксперименты по подбору мет-

рик семантической близости, наиболее полезных при идентификации кореферентных связей. Их результаты показали, что семантические признаки, основанные на подобранных метриках, дают значительно более ощутимое улучшение, чем семантические признаки, описанные в статье С. Толдовой и М. Ионова.

Вторая статья описывает результаты эксперимента по использованию информации из русской Википедии для автоматического распознавания кореферентности. Положительные результаты, полученные на небольшом корпусе с применением признаков, основанных на пересечениях между текстами статей, позволили продолжить эксперименты с данным видом информации, описанные далее.

Эксперименты

Для исследования различных параметров интеграции семантической информации в систему распознавания кореферентных связей была проведена серия экспериментов. Эксперименты были призваны решить следующие задачи: оценить различные метрики семантической близости двух референциальных выражений и выделить метрики, использование которых дает наилучшие результаты, а также оценить эффективность использования энциклопедической информации для обсуждаемой задачи. Данными для экспериментов, описываемых ниже, послужил корпус RuCor, подготовленный в рамках соревнования Ru-Eval-2014.

Использованные данные

Эксперименты проводились на материале корпуса RuCor [Toldova et al., 2016], собранного для соревнования систем автоматического распознавания кореферентности, проводившегося в рамках форума RuEval-2014. В корпус входит 180 текстов различных жанров, содержащих в общей сложности 3 638 кореферентных цепочек и 16 557 кореферентных групп. Все тексты токенизированы, морфологически и синтаксически размечены с помощью инструментов, описанных в [Sharoff, Nivre, 2001]. В соревновании корпус использовался в том числе в качестве «золотого» стандарта разметки, в связи с чем в нем выделены и снабжены соответствующими тегами все кореферентные выражения.

Первый этап экспериментов

Целью первого этапа экспериментов была разработка оптимального представления используемой для анализа семантической информации. Для этого в качестве альтернативного источника информации был проанализирован тезаурус RuThes, и подвергнуты сравнению различные метрики семантической близости между референтами.

Альтернативным источником энциклопедических знаний, использованным на данном этапе, стал тезаурус русского языка RuThes, описанный в работе [Лукашевич, 2011]. Его опубликованная часть RuThes-lite включает в себя 55 тысяч сущностей, которым соответствует 158 тысяч лексических входов, слов и выражений. Сущности внутри онтологии соединены между собой отношениями *выше* / *ниже*, *часть* / *целое*, а также ассоциативными связями.

В качестве альтернативы Википедии был выбран именно этот ресурс, потому что по многим признакам, как структурным, так и содержательным, представлял ее полную противоположность:

1) Википедия является «открытой энциклопедией», т. е. она доступна для редактирования любому пользователю, но ее пополнение не является чьей-либо обязанностью. В связи с этим информация в Википедии часто может быть неполной или устаревшей, а иногда и неверной. RuThes же создавался лингвистами, и он содержит экспертную информацию, которая организована в соответствии с теоретическими принципами построения онтологий и, следовательно, должна быть непротиворечивой, последовательной и корректной;

2) категории Википедии не полностью соответствуют иерархической структуре онтологии, которой является RuThes. Кроме того, они являются намного более дробными по содер-

жанию; страницы Википедии часто могут содержать техническую информацию или метаинформацию, не относящуюся к содержанию статьи, что затрудняет их анализ;

3) в RuThes, в отличие от Википедии, у многих ключей отсутствуют словарные статьи, вместо которых даны только списки синонимов.

В системе распознавания кореферентности использовалась информация о семантической близости двух именных групп. Под семантической близостью понимается то, насколько тесно сущности связаны между собой связями в онтологии. Меры семантической близости успешно используются для снятия семантической неоднозначности, создания вопросно-ответных систем, в том числе для разрешения кореферентности. С одной стороны, эта метрика достаточно прозрачно представляет степень сходства, вплоть до совпадения, между объектами, с другой стороны, она по определению представляется в численном виде, а потому легко может быть интегрирована в автоматические системы в качестве параметра.

Существует большое количество работ, предлагающих методы подсчета семантической близости между сущностями (см. [Budanitsky, Hirst, 2006; Крюков и др., 2010] для описания и сравнения различных метрик). В рамках эксперимента был проведен сравнительный анализ того, насколько меры семантической близости, подсчитанные несколькими различными методами, удовлетворяют задачам автоматического разрешения кореферентности.

Отобранные для сравнения меры разбивались на три больших класса в зависимости от принципа подсчета, на котором они основаны. К первому относились меры, основанные на иерархической структуре онтологии – опирающиеся на расстояние между сущностями. Основной из них является предложенная в статье [Rada et al., 1989] длина пути $path(c_1, c_2)$ между вершинами онтологии c_1 и c_2, соответствующими искомым сущностям (далее в тексте *rada*).

В работе [Wu, Palmer, 1994] предлагается нормализовать данную меру с учетом максимальной глубины онтологии *D*:

$$wp(c_1, c_2) = -\log \frac{path(c_1, c_2)}{2D}.$$

Другой вариант нормализации, с учетом глубины узлов онтологии описан в статье [Leacock, Chodorow, 1998]:

$$lc(c_1, c_2) = \frac{2 * depth(lcs)}{depth(c_1) + depth(c_2)}, \quad (1)$$

где $depth(n)$ – глубина узла *n*; *lcs* – ближайший общий родительский узел узлов c_1 и c_2.

Второй класс составили меры, основанные на подсчете информационной содержательности (information content) понятия. Под информационной содержательностью понимается частота встречаемости понятия, которая подсчитывается как вероятность его появления в корпусе текстов. Термин впервые встречается в статье [Resnik, 1995], метод же подсчета семантической близости на основе этой величины описан в работе [Seco et al., 2004]. Авторы предлагают оценивать близость между узлами онтологии c_1 и c_2 по внутренней информационной содержательности (intrinsic information content) ближайшего к ним общего родительского узла:

$$res(c_1, c_2) = 1 - \frac{\log(hypo(lcs) + 1)}{\log(\max_{wn})},$$

где *lcs* – общий родительский узел (см. (1)); *hypo*(*n*) – число гипонимов узла *n*; max(*wn*) – число узлов онтологии.

Третий класс был представлен мерой текстовых пересечений (text overlaps), предложенной М. Леском [Lesk, 1986] для подсчета близости между словарными определениями сущностей. Модификация этой меры для расширенных текстовых пересечений (extended text

overlaps), предназначенная для подсчета семантической близости между статьями Википедии, описана в статье [Banerjee, Pedersen, 2003].

Ее вычисление основано на количестве пересечений различной длины между текстами, в случае Википедии между первым абзацем статьи и ее полным текстом: $overlap(t_1, t_2) = \sum_n m^2$, для m n-словных пересечений между текстами t_1 и t_2. В статье [Ponzetto, Strube, 2006] предлагается метод нормализации данной метрики с учетом длины сравниваемых текстов:

$$lesk(t_1, t_2) = \tanh\left(\frac{overlap(t_1, t_2)}{length(t_1) + length(t_2)}\right).$$

После того, как были выбраны источники и способ представления данных, было необходимо оценить, насколько полученная информация может быть использована для разрешения кореферентности. Оценка репрезентативности пяти выбранных мер (*rada*, *wp*, *lc*, *res*, *lesk*) была проведена на материале подкорпуса корпуса RuCor [Toldova et al., 2016], созданного в рамках проведения упоминавшегося выше форума RuEval-2014. Корпус RuCor использовался в качестве золотого стандарта для систем, участвовавших в соревновании, в связи с чем в нем вручную выделены все выражения, являющиеся кореферентными, что позволило использовать данные из него, не прибегая к дополнительной разметке. Целью эксперимента было сравнение того, насколько значения мер, полученные для пар сущностей, являющихся или не являющихся кореферентными, коррелируют с разметкой и, следовательно, могут использоваться для автоматического анализа.

Для анализа были взяты кореферентные цепочки из 55 текстов, входивших в тестовую выборку при проведении соревнования RuEval-2014, в которых в общей сложности было выделено 1 016 кореферентных цепочек. В экспериментальный корпус вошло 200 пар кореферентных выражений из цепочек и такое же количество референциальных выражений, не входивших ни в одну кореферентную цепочку. Затем, используя Википедию и RuThes в качестве двух различных источников данных, для каждой пары именных групп мы подсчитали значения пяти описанных выше метрик.

Как нижний порог оценки качества для каждой пары (i, j) была подсчитана мера Жаккара. Она вычислялась на основе количества результатов (*hits*) в выдаче поисковой системы Google по запросам, содержащим элементы пары *i* и *j* по отдельности и вместе:

$$jaccard = \frac{hits(i \ \ and \ \ j)}{hits(i) + hits(j) - hits(i \ \ and \ \ j)}.$$

В табл. 1 приведен пример значений всех метрик, подсчитанных для пары кореферентных и пары некореферентных сущностей.

Наконец, для всех значений каждой из описанных выше метрик, вычисленных на материале обоих источников, был определен коэффициент корреляции Пирсона с корпусной разметкой. Для того чтобы осуществить сравнение было возможно, кореферентные пары условно считались имеющими максимальное значение соответствующей метрики, а некореферентные – минимальное.

В табл. 2 приведены полученные значения коэффициента корреляции для оценивавшихся мер, включая базовое значение коэффициента Жаккара. Меры сгруппированы по авторам статей, в которых они были впервые предложены, и источнику семантических данных. Для Википедии приведены два набора мер: для всех сущностей в выборке и отдельно только для стимулов, включавших в себя именованные сущности. Мера *lesk* не могла быть подсчитана на данных из RuThes, поскольку глоссы в нем приведены лишь для очень небольшого числа сущностей.

Таблица 1

Пример подсчета метрик семантической близости
для пар референциальных выражений

Table 2

Example of calculating semantic relatedness measures for pairs
of referential expressions

Пара	*jaccard*	RuThes					Википедия				
		rada	*wp*	*lc*	*res*	*lesk*	*rada*	*wp*	*lc*	*res*	*lesk*
такса – собака	0,07	2	0,07	0,93	0,8	–	3	1,47	1	0,27	0,025
здание – кабинет	0,06	5	0,03	0,89	0,74	–	5	1,17	0,77	0,73	0,031

Таблица 2

Значения коэффициента корреляции метрик семантической близости
с кореферентной разметкой

Table 2

Values for coefficient of semantic relatedness measures correlation
and coreference annotation

Источник	*jaccard*	*rada*	*wp*	*lc*	*res*	*lesk*
RuThes	0,34	0,56	0,51	0,59	0,30	n/a
Википедия	0,34	0,05	0,58	0,35	0,23	0,03
Википедия (именованные сущности)	0,6	0,7	0,08	0,6	0,2	0,2

Как видно из табл. 2, меры, подсчитанные на основе данных RuThes, в целом оказались более репрезентативными, чем меры, основанные на данных Википедии. Это может быть связано с особенностями структуры дерева категорий онлайн-энциклопедии, которое содержит не только иерархические отношения между сущностями, но и большое количество специальных и метакатегорий. Такая структура может приводить к тому, что путь в дереве между двумя некореферентными сущностями оказывается значительно короче благодаря существованию общей для них метакатегории (например, статьи «Барак Обама» и «Билл Клинтон» будут иметь общую родительскую категорию «Президенты США» притом, что соответствующие сущности явно не являются кореферентными).

Меры, основанные на расстоянии между сущностями, как обычная мера *rada*, так и ее нормализованные варианты, оказались наиболее репрезентативными из рассмотренных. Мера информационной содержательности оказалась менее репрезентативной для данных из обоих источников. Одной из причин этого может являться то, что узлы онтологии более высокого уровня могут иметь меньше непосредственных гипонимов, чем узлы более низкого уровня. Представляется возможным бороться с этим, используя для подсчета метрики количество не только прямых гипонимов узла, но и всех узлов, лежащих ниже в онтологии. Таким образом, для сущностей, общий родительский узел которых лежит ниже в онтологии, значение меры будет выше, чем если такой узел будет находиться ближе к ее вершине.

Мера семантической близости, основанная на текстовых пересечениях, оказалась наименее репрезентативной из всех рассмотренных и не использовалась в дальнейшем. Низкие значения меры, свидетельствующие о небольшом объеме текстовых совпадений между статьями, могут быть связаны с различными особенностями стиля написания. Данный вопрос

выходит за рамки настоящей работы, но представляется интересным для тематических исследований.

Несмотря на более высокую степень корреляции мер, основанных на данных тезауруса, с кореферентной разметкой, из табл. 2 также видно, что меры близости, подсчитанные на данных Википедии отдельно только для тех стимулов, которые содержат именованные сущности, оказываются выше базового значения и вполне сравнимы со значениями мер, подсчитанных на данных RuThes. Этот факт оказывается важен в связи с тем, что в тезаурусе полностью отсутствуют данные о персоналиях, и доступная семантическая информация о них может быть взята только из Википедии. Таким образом, наилучшим способом использовать рассмотренные источники данных было совместить информацию из обоих для подсчета мер семантической близости, используя один из них в зависимости от конкретного типа референта.

Второй этап экспериментов

Заключительным этапом в описываемой в данной работе серии экспериментов стало создание системы автоматического разрешения кореферентности, имплементирующей рассмотренные выше меры семантической близости в качестве одного из параметров. Разработанная система использовала подход, основанный на машинном обучении, поскольку, как правило, именно этот класс систем показывает лучшие результаты на соревнованиях (см. [Pradhan et al, 2012; Toldova et al., 2014]). Система была создана на основе описанной в статье [Kutuzov, Ionov, 2014] модели, использующей в качестве алгоритма классификации деревья решений (decision trees).

Использованная система основана на алгоритме попарной классификации и включает в себя признаки, основанные на расстоянии между членами пары и их морфологическими характеристиками, но кроме того учитывает и синтаксические и простейшие семантические признаки, такие как проверки на то, является ли одна из групп в паре аппозитивной к другой или являются ли члены пары именами собственными. Итоговый набор признаков совпадает с 11 основными признаками, описанными в статье [Toldova, Ionov, 2017], кроме того, были имплементированы дистанционные и морфологические признаки из упомянутой статьи.

Эксперимент, как и предыдущий этап работы, проводился на основе корпуса RuCor, который для этого был разбит на два подкорпуса – обучающий и тестовый (70 и 30 % объема исходного корпуса соответственно). Оценка качества работы системы осуществлялась с помощью набора инструментов, использовавшегося в рамках конференции CONLL-2012 [Pradhan et al., 2012] и считающегося стандартом при анализе кореферентности. Для непосредственной оценки использовались две метрики: MUC [Vilain et al., 1995] и B^3 [Bagga, Baldwin, 1998], являющаяся модифицированной версией первой. В соответствии с каждой метрикой измерялись точность, полнота и F-мера качества работы системы.

В ходе эксперимента были подсчитаны значения метрик качества для следующих версий системы: не использующей семантическую информацию (в табл. 3 *Soon*), опирающейся на семантические признаки только из одного источника и опирающейся на семантические признаки из обоих источников. В табл. 3 приведены значения метрик для всех четырех рассматривавшихся версий системы.

Из приведенных результатов видно, что в то время, как использование признаков на основе Википедии позволяет повысить полноту результатов, использование признаков на основе RuThes увеличивает точность распознавания связей. Таким образом, совмещение признаков, полученных из обоих источников, позволяет достигнуть максимального улучшения качества работы системы. Этим подтверждаются выводы, сделанные на основе второго этапа экспериментов, о полезности информации из Википедии, несмотря на более низкие параметры корреляции.

Таблица 3

Метрики качества для различных версий системы

Table 3

Quality metrics for different versions of the system

Система	MUC			B^3		
	Точность	Полнота	F-мера	Точность	Полнота	F-мера
Soon	72,76	59,49	65,46	71,01	44,50	54,71
Soon + Википедия	70,28	59,71	64,56	66,50	44,63	53,41
Soon + RuThes	72,72	59,43	65,41	71,15	44,44	54,71
Soon + RuThes + Википедия	73,57	60,01	**66**,**10**	71,77	44,93	**55**,**26**

Использование семантических признаков из обоих источников позволило улучшить распознавание кореферентности между именными группами, находящимися в нескольких классах отношений. С одной стороны, это таксономические связи, такие как гиперонимия или синонимия, в том числе между именованными сущностями (примеры (2)–(4)).

(2) Сразу же с момента издания [**альбом**] стал настолько популярным, <…> что звукозаписывающая компания EMI не успевала штамповать копии [**пластинки**] для всех желающих ее приобрести.

(3) Выжившие после крушения [**корабля**] рассказали, что основная причина трагедии – то, что [**теплоход**] был очень старый.

(4) Альбом первоначально был издан в 1975 году компаниями Harvest Records в [**Великобритании**] и Columbia Records в США. <…> в 1980 году альбом появился в [**Соединённом Королевстве**] с более высоким качеством…

С другой стороны, это отношения между именованной сущностью и именной группой, обозначающей один из аспектов этой сущности (пример (5)).

(5) Виктор Вексельберг хотел бы ангажировать [**Григория Перельмана**] для работы в Кремниевой долине. Фортуна повернулась к [**питерскому математику**] лицом: сначала мировое признание, затем миллион долларов.

Продемонстрированный системой рост качества превышает результат, полученный в аналогичном эксперименте, описанном в статье [Toldova, Ionov, 2017]: максимальный прирост в 0,54 % MUC и 0,55 % B^3 по сравнению с 0,26 и 0,19 % соответственно. Как можно заметить, наибольшее увеличение качества в обоих экспериментах дает использование информации
из RuThes. Поскольку в данном исследовании семантическая информация представлена в виде мер семантической близости в отличие от простого расстояния в онтологии, можно заключить, что более тщательная обработка семантической информации, в частности в форме подобных метрик, является более выигрышным способом представления семантической информации и позволяет в большей степени улучшить работу системы.

Поскольку описанная выше система использует для распознавания кореферентных связей тезаурусную и энциклопедическую информацию, за пределами ее возможностей на данный момент остаются случаи контекстной синонимии, опирающиеся на содержание текста, подобно примеру (6). Для таких случаев представляется необходимым использовать другие признаки, основанные на семантическом анализе текста или анализе совместной встречаемости слов. Подбор и оценка эффективности этих признаков, однако, находятся за рамками данной работы и должны стать предметом отдельного исследования.

(6) Выходец из [**Нигерии**] решил остаться на ПМЖ в Израиле, поскольку на [**родине**] его якобы преследует опасный призрак.

Заключение

В данной работе описан ряд прикладных лингвистических экспериментов, целью которых было, во-первых, продемонстрировать важность семантической информации из энциклопедических источников для разрешения кореферентности и, во-вторых, создать работающую автоматическую систему для данной задачи с интегрированной тем или иным способом семантической информацией.

В ходе первых экспериментов было проведено сравнение различных мер семантической близости между референциальными выражениями, полученных на основе данных Википедии и тезауруса RuThes, с целью сравнения их пригодности для использования в системах распознавания кореферентности. Полученные результаты показали бóльшую релевантность данных RuThes в целом, в то время как данные Википедии оказалось возможным использовать для улучшения разрешения кореферентных связей именованных сущностей. Вторая серия экспериментов была нацелена на оценку качественного улучшения работы системы после имплементации разработанных ранее признаков. В результате разработана работающая система автоматического анализа кореферентности, включающая в себя семантическую информацию. Качество ее работы оказалось выше, чем у аналогичных систем, не использующих семантические признаки и использующих их более простые представления. Таким образом, подтверждается актуальность результатов предыдущего этапа экспериментов.

Важность результатов, полученных в результате проведенной работы, состоит в следующем:

1) экспериментально подтверждено повышение качества автоматического распознавания кореферентности в русском языке при использовании семантической информации из внешних источников;

2) проведены подсчет и сравнение мер семантической близости для русского сегмента Википедии и тезауруса RuThes, что позволило определить более репрезентативные меры такого рода и указало на ряд особенностей структуры Википедии, негативно влияющих на результаты их вычисления;

3) оценен вклад мер семантической близости в качество распознавания кореферентности в сравнении с другими признаками, основанными на семантической информации;

4) создана работающая система автоматического разрешения кореферентности с интегрированными семантическими признаками, доступная для изучения и улучшения.

Данная работа открывает несколько различных направлений дальнейших исследований. С одной стороны, существует большой объем задач по более тщательной оценке мер семантической близости для онтологий на русском языке. Существует также большое количество не рассмотренных в данной работе метрик, в том числе специализированных алгоритмов для оценки близости сущностей в Википедии (в частности, алгоритм, описанный в [Yeh et al., 2009]). Кроме того, описанные результаты могут открыть путь к дальнейшему изучению способов интеграции семантической информации в системы автоматического распознавания кореферентности как с точки зрения альтернативных источников данных, так и с точки зрения альтернативных способов их представления. Необходимо также учесть существование случаев контекстной синонимии, которые не могут быть распознаны за счет энциклопедической и тезаурусной информации и требуют иного подхода к анализу.

Наконец, интерес для теоретических исследований могут представлять затронутые в данной статье особенности структуры русской Википедии как онлайн-энциклопедии и ее возможные различия по сравнению как с традиционными энциклопедиями, так и с другими источниками структурированных данных: тезаурусами, такими как RuThes, или иноязычными сегментами Википедии.

Список литературы / References

Крюков К. В., Панкова Л. А., Пронина В. А., Суховеров В. С., Шипилина Л. Б. Меры семантической близости в онтологии // Проблемы управления. 2010. Вып. 5. С. 2–14.
Kryukov, K. V., Pankova, L. A., Pronina, V. A., Sukhoverov, V. S., Shipilina, L. B. Semantic similarity measures in ontology. Review and classification. *Control Sciences*, 2010, iss. 5, p. 2–14. (in Russ.)

Лукашевич Н. В. Тезаурусы в задачах информационного поиска. М.: Изд-во Моск. ун-та, 2011.
Loukachevitch, N. Thesauri in Information retrieval tasks. Moscow, Moscow State University Publ., 2011. (in Russ.)

Толпегин П. В. Автоматическое разрешение кореференции местоимений третьего лица русскоязычных текстов: Дис. ... канд. техн. наук. М., 2008.
Tolpegin, P. V. Automated resolution of third person pronouns coreference in Russian texts. PhD dissertation. Moscow, 2008. (in Russ.)

Azerkovich, I. Employing Wikipedia data for coreference resolution in Russian. *Artificial Intelligence and Natural Language. AINL 2017. Series: Communications in Computer and Information Science*, 2018, vol. 789, p. 107–112.

Bagga, A., Baldwin, B. Algorithms for scoring coreference chains. In: The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference, 1998, vol. 1, p. 563–566.

Banerjee, S., Pedersen, T. Extended gloss overlaps as a measure of semantic relatedness. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc., 2003, p. 805–810.

Bogdanov, A. V., Dzhumaev, S. S., Skorinkin, D. A., Starostin, A. S. Anaphora analysis based on ABBYY Compreno linguistic technologies. In: Computational linguistics and intellectual technologies: Proceedings of the international conference "Dialogue 2014". Moscow, 2014, vol. 13 (20), p. 89–102.

Budanitsky, A., Hirst, G. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 2006, vol. 32 (1), p. 13–47.

Haghighi, A., Klein, D. Simple coreference resolution with rich syntactic and semantic features. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2009, vol. 3, p. 1152–1161.

Harabagiu, S. M., Bunescu, R. C., Maiorano, S. J. Text and knowledge mining for coreference resolution. In: Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. Association for Computational Linguistics, 2001, p. 1–8.

Kamenskaya, M. A., Khramoin, I. V., Smirnov, I. V. Data driven methods for anaphora resolution of Russian texts. Computational linguistics and intellectual technologies: Proceedings of the international conference "Dialogue 2014". Moscow, 2014, vol. 13 (20), p. 241–250.

Kameyama, M. Recognizing referential links: an information extraction perspective. In: Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution. Madrid, Spain, 1997, p. 46–53.

Kutuzov, A. B., Ionov M. The impact of morphology processing quality on automated anaphora resolution for Russian. In: Computational linguistics and intellectual technologies: Proceedings of the international conference "Dialogue 2014". Moscow, 2014, vol. 13 (20), p. 232–240.

Leacock, C., Chodorow, M. Combining local context and WordNet similarity for word sense identification. In: WordNet. An Electronic Lexical Database. MIT Press, 1998, p. 265–283.

Lesk, M. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: Proceedings of the 5th Annual Conference on Systems Documentation. ACM, 1986, p. 24–26.

Ng, V., Cardie, C. Improving machine learning approaches to coreference resolution. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002, p. 104–111.

Ponzetto, S. P., Strube, M. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Association for Computational Linguistics, 2006, p. 192–199.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., Zhang, Y. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012, p. 1–40.

Rada, R., Mili, H., Bicknell, E., Blettner, M. Development and application of a metric to semantic nets. In: IEEE Transactions on Systems, Man and Cybernetics, 1989, iss. 19 (1), p. 17–30.

Rahman, A., Ng, V. Coreference resolution with world knowledge. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011, vol. 1, p. 814–824.

Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the Fourteenth International. Joint Conference on Artificial Intelligence. IJCAII, 1995, vol. 1, p. 448–453.

Seco, N., Veale, T., Hayes, J. An intrinsic information content metric for semantic similarity in WordNet. In: Proceedings of the 16th European Conference on Artificial Intelligence. IOS Press, 2004, p. 1089–1090.

Sharoff, S., Nivre, J. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In: Computational linguistics and intellectual technologies: Proceedings of the international conference “Dialogue 2014”. Moscow, 2011, vol. 10 (17), p. 591–605.

Soon, W. M., Ng, H. T., Lim, D. C. Y. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 2001, vol. 27, no. 4, p. 521–544.

Toldova, S. Ju. Roytberg, A., Ladygina, A. A., Vasilyeva, M. D., Azerkovich, I. L., Kurzukov, M., Sim, G., Gorshkov, D. V., Ivanova, A., Nedoluzhko, A., Grishina, Y. Ru-Eval-2014: Evaluating anaphora and coreference resolution for Russian. In: Computational linguistics and intellectual technologies: Proceedings of the international conference “Dialogue 2014”. Moscow, 2014, vol. 13 (20), p. 681–694.

Toldova, S., Grishina, Yu., Ladygina, A., Vasilyeva, M., Sim, G., Azerkovich, I. Russian coreference corpus. In: Almeida F. A., Barrera I.O., Toledo E. Q. (eds.). Input a Word, Analyze the World: Selected Approaches to Corpus Linguistics. Cambridge Scholars Publishing, 2016, p. 107–124.

Toldova, S., Ionov, M. Coreference Resolution for Russian: The Impact of Semantic Features. In: Computational linguistics and intellectual technologies: Proceedings of the international conference “Dialogue 2014”. Moscow, 2017, vol. 16 (23), p. 339–348.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L. A model-theoretic coreference scoring scheme. In: Proceedings of the 6th Conference on Message Understanding, ser. MUC6 ’95. Stroudsburg, PA, USA: Association for Computational Linguistics, 1995, p. 45–52.

Wu, Z., Palmer, M. Verb semantics and lexical selection. In: Proceedings of ACL-94, 1994, p. 133–138.

Yeh, E., Ramage, D., Manning, C. D., Agirre, E., Soroa, A. WikiWalk: random walks on Wikipedia for semantic relatedness. In: Proceedings of the 2009 workshop on graph-based methods for natural language processing. Association for Computational Linguistics, 2009, p. 41–49.

Материал поступил в редколлегию
Date of submission
08.09.2018

Сведения об авторе / Information about the Author

Азеркович Илья Леонидович, аспирант аспирантской школы по филологическим наукам факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики» (ул. Ст. Басманная, 21/4, Москва, 105066, Россия)

Ilya L. Azerkovich, National Research University Higher School of Economics (21/4 Staraya Basmannaya Str., Moscow, 105066, Russian Federation)

iazerkovich@gmail.com
SPIN-код 7211-6400
ORCID 0000-0002-6482-1137