О влиянии семантики на точность определения парафраз в русскоязычных текстах

© К.К. Боярский[1] © Е.А. Каневский[2]

[1] Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики,
[2]Санкт-Петербургский экономико-математический институт РАН,
Санкт-Петербург

boyarin9@yandex.ru kanev@emi.nw.ru

Аннотация. Статья посвящена идентификации парафраз в русскоязычных текстах. В качестве инструмента для решения этой проблемы предлагается использовать семантико-синтаксический парсер SemSin и семантический классификатор. Проанализированы несколько вариантов определения близости пар предложений: по словам, по леммам, по классам, по семантически связанным концептам, по предикатным группам. Обсуждены преимущества и недостатки этих методов. Показано, что при увеличении глубины использования семантической информации качество идентификации парафраз повышается. Однако включение в анализ предикатных групп, определяемых по дереву зависимостей, может даже привести к ухудшению качества идентификации вследствие увеличения числа ложноположительных решений.

Ключевые слова: парафразы, семантические словари, леммы, классификатор, семантические классы, парсинг, синонимия.

Effect of Semantic Parsing Depth on the Identification of Paraphrases in Russian Texts

© K. Boyarsky[1] © E. Kanevsky[2]

[1]ITMO University,
[2]EMI RAS,
St Petersburg, Russia

boyarin9@yandex.ru kanev@emi.nw.ru

Abstract. As a tool to solve the problem of identification of paraphrases in Russian texts, the paper proposes the semantic-syntactic parser SemSin and a semantic classifier. Several alternative methods for evaluating the similarity of sentence pairs – by words, by lemmas, by classes, by semantically related concepts, by predicate groups – have been analyzed. Advantages and drawbacks of the methods are discussed. The paraphrase identification quality has been shown to rise with increasing depth of using the semantic information. Yet, complementing the analysis with predicate groups, identified by the dependency tree, may even cause the identification to degrade due to the growing number of false positive decisions.

Keywords: Russian texts, paraphrases, semantic dictionary, lemmas, classifier, classes, semantic-syntactic parsing, synonymy.

1 Введение

В последнее время значительный интерес исследователей, работающих в области поиска информации, привлекает проблема выявление парафраз.

В англоязычной литературе существует большое количество работ по идентификации парафраз с привлечением различной лексической, синтаксической и семантической техники [3, 17]. В большинстве способов использовалось обучение, проводились токенизация, определение частей речи и обработка только существительных и глаголов [6]. Использовалось также придание различного веса словам с учетом их грамматической роли в предложениях. Corley and Mihalcea [4] использовали измерения семантической близости текстов с помощью WordNet [8]. При этом семантическое сходство слов измерялась только для глаголов и существительных, а сравнение наречий, прилагательных и количественных числительных

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

проводилось лексически. Было показано, что такой метод значительно точнее, чем простое лексическое сравнение. Близкая техника, основанная на семантической информации WordNet, использовалась в [5]. Pershina [13] дополнительно для идентификации парафраз использовала базу идиом. Лучшие результаты по идентификации парафраз в англоязычных текстах дают F-меру около 82%.

Для русскоязычных текстов работы по идентификации парафраз весьма немногочисленны [18]. Значительные сложности возникают в связи с особенностями русского языка, который отличается свободным порядком слов и богатой морфологией [16]. Целью настоящей работы являются исследование эффективности различных вариантов анализа парафраз в русскоязычных текстах и определение оптимальной степени использования семантической информации.

Данная работа проводилась в рамках конкурса по идентификации русских парафраз, представленных в корпусе paraphraser.ru [10]. В рамках этого конкурса предлагалось два варианта анализа пар новостных заголовков: разделение на две группы (парафразы и не парафразы) и на три группы, с выделением дополнительно группы нечетких парафраз. По представленным данным оказалось, что первый вариант анализа обладает существенно большей точностью. В значительной мере это связано с субъективностью выделения нечетких парафраз. Поэтому было выбрано разделение на две группы.

Для определения того, являются ли два предложения парафразами, т. е. одинаков ли их смысл для носителей языка, необходимо ввести числовую меру сходства. В качестве такой меры мы использовали коэффициент Жаккара J. Если два предложения A и B содержат соответственно $n(A)$ и $n(B)$ лексических единиц, то

$$J = \frac{n(A) \cap n(B)}{n(A) \cup n(B)}.$$

В соответствии с этим критерием мера сходства определяется как отношение числа совпадающих единиц к общему числу различных единиц.

Следует решить два вопроса: что считать лексической единицей, подлежащей сравнению, и каков критерий собственно сравнения. Для сравнения нами были выбраны четыре варианта.

Вариант 1. В качестве единицы для сравнения принимается слово или иная непрерывная последовательность знаков. Предполагается, что эта последовательность достаточно длинная, так что одно- и двухбуквенные слова не учитываются.

Преимущества. Не требуется каких-либо средств для разбора текста. Выявление совпадающей группы знаков с большой вероятностью означает наличие однокоренных слов, близких по смыслу. Если же такая группа представляет отдельное слово, то возможно, что это слово выполняет одну и ту же функцию в обоих высказываниях.

Недостатки. Похожие фрагменты с большой вероятностью включают в себя служебные слова. Их совпадение часто носит случайный характер и не отражает смысла фрагмента текста (хотя, например, совпадение отрицания *не* может служить для определения меры близости). Кроме того, поскольку русский язык относится к синтетическому типу, то даже очень небольшая перефразировка, совершенно не меняющая смысл высказывания, приводит к изменению словоформ. Это существенно снижает точность анализа.

Вариант 2. В качестве единицы для сравнения вместо словоформы принимается нормализованная форма слова (лемма). Исключаются служебные слова. Производится настройка парсера на предметную область путем исключения омонимичных словоформ, не принадлежащих к данной предметной области. Например, словоформе *белку* соответствуют две леммы: *белок* и *белка*. В тексте по биологии будет выбрана только первая из них.

Преимущества. Устраняются сложности, связанные с богатством словоформ каждой леммы. Повышается независимость анализа от конкретного строя предложения (*человек, который смеется* vs *смеющийся человек*).

Недостатки. Требуется инструмент для морфологического анализа. Достаточно часто (примерно в 7% случаев) лемма по словоформе определяется неоднозначно, т. е. возникает проблема омонимии. Остаются вопросы, связанные со служебными словами.

Вариант 3. При сравнении учитывается семантика слов.

Преимущества. Слова сравниваются не столько по написанию, сколько по смыслу, что, в принципе, должно повысить точность определения парафраз.

Недостатки. Сложность и неоднозначность семантического анализа. Необходимость использования семантических словарей. Если в английском языке эту функцию выполняет semantic web, то для русского языка соответствующие инструменты слабо развиты, отсутствует принятый в качестве стандарта де-факто семантический словарь.

Вариант 4. Сравнение производится с учетом полного дерева разбора.

Преимущества. Возможность анализа сходства неконтактно стоящих групп слов, выделения смысловых блоков, описывающих термины предметной области.

Недостатки. Повышение вероятности ошибок парсера. Недостаточная проработанность вопроса о выделении контекстных блоков. Чувствительность метода к замене одних оборотов другими, например, причастного оборота придаточным предложением.

Из русскоязычных парсеров, способных проводить глубинный анализ предложения, наиболее известны ЭТАП-3 [12], а также парсеры фирм Яndex [2], Abbyy [1], «Диктум» [9]. Два первых парсера работают с помощью системы правил, два других используют свою собственную технологию. Все парсеры тем или иным образом используют словари.

В данной работе использовался семантико-синтаксический парсер SemSin [22]. Это система, сочетающая в себе функции лемматизатора,

синтаксического и семантического анализатора. Парсер включает в свой состав словари, классификатор, блок морфологического анализа, предсинтаксический модуль [21] и набор продукционных правил [23,24].

В процессе анализа предложения одновременно выполняются снятие грамматической и частеречной омонимии, сегментация предложения и построение синтаксического дерева зависимостей.

Полученное дерево содержит максимально полную информацию о предложении. Эта информация может в дальнейшем служить основой для решения самых разных задач: выявления терминов [15], классификации текстов [7] и т. д. В данной работе обсуждается, какая именно информация полезна для определения близости смысла предложений, т. е. для идентификации парафраз.

2 Анализ текста

Основное внимание было уделено анализу различных способов описания семантики и включения ее в процедуру идентификации парафраз. Все примеры и экспертные оценки брались из корпуса русских парафраз [19].

2.1 Лемматизация

Как было показано в [14], в русском языке (в отличие от английского и французского) в задачах кластеризации текстов наибольшей дифференцирующей силой обладают существительные. Однако для идентификации парафраз выделение только существительных слишком грубо и не может охватить всех тонкостей смысла. Поэтому для сравнения оставлялись существительные, прилагательные, глаголы и отглагольные формы (причастия, деепричастия) и числительные. Примеры влияния лемматизации на величину коэффициента Жаккара приведены в табл. 1.

Таблица 1 Влияние лемматизации

	Предложения	Словоформы	Леммы
1	NI опубликовал список самого опасного вооружения флота России В США опубликован топ-5 самых опасных вооружений ВМФ России	0.067	0.455
2	Путин впервые объявил минуту молчания на параде Победы Пан Ги Мун поблагодарил Путина за организацию Парада Победы	0.067	0.250

Отметим, что лемматизация увеличивает степень согласованности предложений как в случае, когда они являются парафразами (пример 1), так и когда совпадение слов случайно, а совпадения смысла нет (пример 2). Таким образом, лемматизация безусловно повышает точность сравнения лексики предложений, но недостаточна для сравнения смысла.

2.2 Семантика. Учет классов

Для более точного сравнения предложений был использован семантический классификатор, содержащий около 1700 классов. Его основой является классификатор Тузова [27], ориентированный именно на компьютерный анализ текстов. Дерево классов построено с таким расчетом, чтобы семантические классы имели определенные синтаксические свойства. Например, список действий, которые может производить живое существо, отличается от действий неодушевленных предметов, у разных классов могут быть разные атрибуты и т. д. Кроме того, формат этого классификатора удобен для его автоматического использования.

Наш классификатор отличается, в частности, от классификатора Шведовой [26]. Например, слово *жрец* в обоих случаях классифицируется практически одинаково: как определенный тип профессии человека. В то же время слово *желание* у Шведовой находится в ветви дерева *духовный мир–чувства–*…, в то время как в нашем классификаторе психические явления и чувства рассматриваются как свойства человека. В классификаторе WordNet положение слова *priest* в общем соответствует русским классификаторам, слова *desire* – ближе к классификатору Шведовой.

При разработке классификатора очень трудно, если вообще возможно, определить, на каком ярусе дерева нужно поместить то или иное слово, по какому признаку разделять подклассы, в какой момент прекращать дальнейшее ветвление. Например, *коса* (*scythe*) в WordNet относится к классу режущих инструментов, а в нашем классификаторе и у Шведовой – к классу сельскохозяйственных орудий. При идентификации парафраз принималось, что принадлежность разных слов в первом и втором предложениях одному классу означает близость их смысла.

Часто вместо конкретного слова в тексте появляется его гипероним. Для обнаружения гиперонимов класс определялся с точностью ±1 уровень иерархии. Таким образом обеспечивается совпадение слов, выделенных полужирным шрифтом в следующих примерах.

Лавров подарил Керри ***помидоры****… vs Лавров подарил Керри* ***овощи****…*

Жертвами взрыва … стали не менее трех ***человек*** *vs Жертвой взрыва… стал* ***гражданин*** *Великобритании.*

Исключение составляют имена собственные, поскольку, например, все названия городов относятся к одному классу, также к одному классу относятся все фамилии людей.

Сравнением классов во многих случаях перекрываются отношения синонимии. Однако следует иметь в виду, что из-за морфологических

особенностей в русском языке значительно меньше совпадений словоформ у различных частей речи. Так? в английском *gold* это и существительное, и прилагательное, в русском – существительное *золото*, прилагательное *золотой*. При определении синонимии в русскоязычном RusNet [20] подразумевается совпадение частей речи у синонимичных слов. Сравнение «по классам» в этом смысле шире и позволяет делать заключение о близости смысла даже при существенной перефразировке:

Турецкий сухогруз подвергся ***обстрелу****… vs Турецкий сухогруз* ***обстреляли****.*

Несомненно, вопрос о том, являются ли два существительных, относящихся к одному классу, синонимами, неоднозначен. Например, слова *веревка*, *бечевка* легко могут взаимозаменяться в тексте, а слова *помидор*, *огурец* – нет. Тем не менее, анализ показывает, что близость классов чаще всего означает близость смысла.

2.3 Семантика. Синонимия и семантические гнезда

В ряде случаев сравнения по классам оказывается недостаточно. Известна, например, «теннисная проблема» [25], заключающаяся в том, что слова, относящиеся к одной предметной области, часто находятся в совершенно разных ветвях классификатора, что затрудняет не только идентификацию парафраз, но и решение задач классификации и кластеризации текстов. Наш словарь содержит дополнительную информацию о семантической близости слов. Будем называть группу семантически связанных слов семантическим гнездом. Фактически появление дополнительных связей означает превращение дерева классов в семантическую сеть. Имеется несколько ситуаций, приводящих к образованию семантических гнезд.

- Производные слова

Все лексемы в словаре делятся на базовые и производные, причем под производными подразумеваются не только слова, однокоренные с базовыми [27]. Базовых лексем несколько меньше половины (около 83000). Из них примерно 20000 образуют семантические гнезда, к которым относятся производные слова с близким смыслом. Так, к гнезду базового слова *чувство* принадлежит более 100 производных слов, среди которых есть существительные (*нечувствительность*, *аналгезия*), прилагательные (*чувствительный*, *душещипательный*), глаголы (*чувствовать*, *обуревать*). В некоторых случаях семантический класс производного слова совпадает с классом базового, в других – отличается. Например, базовое слово *сигнал* относится к подклассу ветви семантического дерева *информация*. Производные от него слова: *сигнальный, сигнализация, сигнализировать* имеют тот же класс, а слово *сигнальщик* – человек с определенным родом занятий – совсем другой. Принадлежность производных слов к общему гнезду добавляет около 7500 связей в семантическую сеть (помимо связей класс – подкласс).

- Географические названия

Как указывалось выше, имена собственные сравниваются только по леммам. Но одна и та же страна может называться по-разному, и это необходимо учитывать при анализе парафраз. Часто в качестве названия страны используется аббревиатура. Поэтому в словарь были внесены дополнительные сведения о тождественной синонимии слов и выражений.

В ***Британии*** *палата общин одобрила однополые браки.*

Палата общин ***Великобритании*** *одобрила однополые браки.*

США просят ***РФ*** *немедленно отменить запрет на ввоз мяса.*

США призвали ***Россию*** *немедленно снять запрет на импорт мяса.*

КНДР *готовится нанести ракетный удар по США.*

Северная Корея *пригрозила ракетным ударом по США.*

Неким аналогом отношений класс – подкласс для географических названий является информация о принадлежности населенного пункта к региону и стране. Такая информация была добавлена в словарь, хотя и в довольно ограниченном размере. Например, указано, что *Дамаск* – столица *Сирии*, а город *Сент-Луис* находится в штате *Миссури* страны *США*:

Неизвестный открыл огонь в бизнес-школе штата ***Миссури****.*

В ***Сент-Луисе*** *преступник открыл огонь в бизнес-школе.*

- Характеристики людей и организаций

Следующей группой слов, для которых в словарь были внесены дополнительные связи, приводящие к образованию семантических гнезд, являются характеристики людей по национальности и месту жительства. Необходимо не только знать, что *сибиряк* – название человека по месту жительства, но и что это место – именно *Сибирь*. Таким образом, каждое такое слово «прикреплено» к соответствующей стране, региону или городу:

Американец *выиграл в лотерею за 100 евро картину П. Пикассо.*

Житель ***Пенсильвании*** *выиграл в лотерею картину Пикассо.*

Для некоторых часто встречающихся имен высокопоставленных деятелей в качестве составляющих семантического гнезда указаны их должности и страны:

Синдзо ***Абэ*** *в письме* ***президенту РФ*** *объяснил, почему не приедет на 9 мая.*

В письме ***Путину японский премьер*** *объяснил причины отказа приехать в Москву 9 Мая.*

США *приостановили поставку истребителей в Египет.*

Обама *приостановил поставки истребителей F-16 в Египет.*

Еще одну группу семантических гнезд составляют связи по принадлежности определенных социальных групп к организациям и т. д. *Коммунист* – наименование человека не просто по принадлежности к какой-то общественной организации, а именно к компартии, а *хоккеист* среди всех видов спорта имеет отношение только к хоккею:

Депутаты от ***КПРФ*** *попросили Путина взять под защиту гималайского медведя.*

Коммунисты *попросили Путина защитить гималайских медведей.*

Сборная России по ***хоккею*** *проиграла финнам и во втором матче Евротура.*

Российские ***хоккеисты*** *проиграли на Евротуре четыре матча подряд.*

- Идиомы

Отдельную группу семантических гнезд образуют связи устойчивых выражений и идиом с их семантическими аналогами:

В Красноярском крае ***исчез*** *заместитель прокурора.*

В Красноярском крае ***пропал без вести*** *помощник прокурора района.*

Ушел из жизни *Уго Чавес.*

Умер *Уго Чавес.*

Рассмотрим следующий пример.

Оппозиция ФРГ угрожает правительству судом из-за шпионского скандала.

Немецкая оппозиция пригрозила правительству иском из-за скандала с BND.

Если сравнивать эти предложения только по леммам, то имеются три совпадения (*оппозиция*, *правительство*, *скандал*) с коэффициентом Жаккара *J*=0.27. При учете классов получаем совпадение для глаголов угрожать и пригрозить, становится *J*=0.40. Название *ФРГ* входит в семантическое гнездо слова *Германия*, в это же гнездо входит слово *немецкий* (производное от базового слова *немец*, которое, в свою очередь связано со словом *Германия*). Получаем *J*=0.55. Наконец, лексемы *суд* и *иск* тоже входят в общее гнездо. Окончательно получаем *J*=0.75, что вполне отражает смысловую близость этих предложений. На данный момент число семантических связей таких типов в словаре около 12 тыс.

2.4 Дерево зависимостей

Нами была сделана попытка использовать построенное парсером дерево зависимостей для уточнения сравнения смысла предложений. Было выдвинуто предположение, что совпадение субъекта действия и собственно предиката повышает вероятность того, что рассматриваемые предложения являются парафразами. В этом случае при расчете коэффициента Жаккара совпадение лемм учитывалось с весовым коэффициентом 1.5.

Однако на практике оказалось, что у пар предложений, обладающих указанным свойством, коэффициент Жаккара обычно и так уже велик. Поэтому его незначительное увеличение сравнительно редко переводит эту пару из разряда «не парафраза» в разряд «парафраза». Но эти редкие переходы тоже представляют определенный интерес. В табл. 2 приведены значения коэффициентов Жаккара при учете совпадений лемм, классов и семантических гнезд (LCS) и дополнительно предикатных пар (LCSP).

Таблица 2 Влияние совпадения пары субъект–предикат на величину коэффициента Жаккара

	Предложения	LCS	LCSP	Комментарий
1	МИД Чехии: дипломат получил выговор за высказывание о пожаре в Одессе Оправдавший сожжение людей в Одессе чешский дипломат получил выговор	0.33	0.50	Верный результат, предложения идентичны по смыслу
2	Кобзон назвал результат России на Евровидении очень достойным Иосиф Кобзон назвал второе место Полины Гагариной достойным	0.27	0.45	Верный результат, предложения идентичны по смыслу
3	Лужков назвал свою ферму примером для российского правительства Лужков назвал российскую экономику «антинародной»	0.37	0.62	Ошибочный рост коэффициента совпадения, связанный с игнорированием прямого дополнения к переходному глаголу
4	МВД насчитало 200 тыс. участников празднования Дня Победы в Севастополе МВД насчитало 250 тыс. участников акции «Бессмертный полк» в Москве	0.28	0.48	Ошибочный рост коэффициента совпадения, связанный с игнорированием различия в месте действия
6	Эксперты из России и Белоруссии направились с проверкой в Эстонию Российские военные эксперты направились с проверкой в Эстонию	0.71	0.86	В стандарте не считаются парафразами. Несомненно, это одно и то же событие, очевидно, ошибка разметки

Таблица показывает, что учет совпадений субъект – предикат в некоторых случаях облегчает нахождение парафраз (примеры 1, 2). Однако это может привести и к ложному повышению степени сходства предложений (примеры 3, 4). В некоторых случаях определение того, является ли пара предложений парафразами, имеет пограничный характер и может интерпретироваться различными экспертами по-разному (пример 5). Иногда возможны просто экспертные ошибки (пример 6). В общем учет совпадений пары субъект – предикат привел к незначительному снижению качества анализа. Однако это изменение лежит в пределах погрешности. Нам представляется перспективным направлением развития работы дальнейшее расширения анализа дерева зависимостей с целью снижения коэффициента сходства у пар предложений, отличающихся, например, по месту действия, и уменьшению числа ложных срабатываний.

3 Результаты

Ниже приведены результаты анализа смысловой близости пар предложений для выявления парафраз, выполненного по следующим схемам.

По словам (W). При сравнении учитывались все слова, включая служебные части речи.

По леммам (L). Учитывались существительные, прилагательные, глаголы, числительные.

По леммам и классам (LC). Дополнительно учитывалось совпадение классов по семантическому дереву.

По леммам, классам и семантическим гнездам (LCS). Дополнительно учитывалось совпадение классов по семантической сети.

Примеры влияния схемы расчета на величину коэффициента Жаккара приведены в табл. 3.

Таблица 3 Коэффициент Жаккара для различных схем подсчета совпадений

№	Предложения	W	L	LC	LCS
1	Крупный пожар вспыхнул на складе на северо-востоке Москвы Крупный пожар в административном здании в центре Москвы потушен	0.230	0.300	0.444	0.444
2	Продажи АвтоВАЗа в России в апреле сократились на 38,3% Продажи «АвтоВАЗа» в России рухнули на треть	0.273	0.500	0.667	0.778
3	СМИ: поздравляя Вакарчука, Кличко ошибся с возрастом и именем юбиляра Кличко перепутал в поздравлении имя и возраст солиста «Океана Эльзы»	0.062	0.230	0.455	0.455
4	Выселен последний экс-депутат, незаконно занимавший жилье в Москве В Москве выселили бывшего депутата Думы из служебной квартиры	0.071	0.181	0.300	0.600
5	Morgan Stanley взял на работу бывшего зампреда Банка России. Бывший глава ФСФР нашел работу в Morgan Stanley	0.200	0.364	0.500	0.500

На этапе обучения системы проводилось пополнение словаря специфическими терминами, а также определение оптимального «уровня отсечки»: с какого значения коэффициента Жаккара считать пару предложений парафразом. Рассчитывались стандартные параметры: аккуратность (A), точность (P), полнота (R) и F-мера (F):

Типичные распределения показаны на рис. 1.

Зависимости имели качественно сходный характер и для остальных схем подсчета. Очевидно, что, снижая уровень отсечки, т. е. относя к парафразам почти все пары предложений, можно добиться сколь угодно высокого значения полноты, а повышая этот уровень, – сколь угодно высокого значения точности. Оценка качества анализа проводилась по параметрам аккуратность и F-мера, для которых оптимальный уровень отсечки оказался в интервале 0.35…0.40.

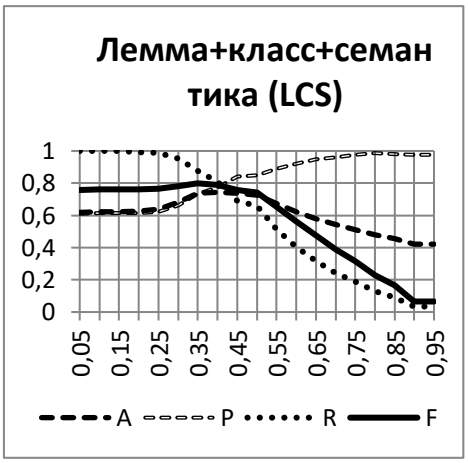

Рисунок 1 Зависимость параметров точности от уровня отсечки

Таблица 4 Аккуратность и F-мера при различных схемах сравнения

Схема	W	L	LC	LCS	LCSP
A	0.692	0.718	0.743	0.744	0.742
F	0.762	0.774	0.784	0.800	0.795

На основании этой таблицы можно сделать следующие выводы.

4 Заключение

Результаты по качеству идентификации парафраз приведены в табл. 4. Сравнение проводилось по разметке Золотого стандарта [11]. Для каждой из рассмотренных схем сравнения приведены лучшие значения аккуратности A и F-меры F.

В большинстве случаев заголовки новостей, относящиеся к одному и тому же событию, лексически очень сходны, и могут быть определены как парафразы любым способом.

Применение методов семантико-синтаксического сравнения (LC) улучшает результаты по сравнению не только с простым посимвольным сравнением, но и со сравнением по леммам.

Увеличение «глубины» семантического анализа за счет перехода от дерева классов к семантической сети (LCS) улучшает качество анализа.

Дополнительный учет совпадений субъект – предикат (LCSP) незначительно ухудшает качество за счет увеличения числа ложных срабатываний.

Заметим, что в статье [18], табл. V, приводятся 15 пар предложений, особенно трудных для анализа парафраз. Три метода, рассматриваемые в этой статье, дают соответственно 6, 6 и 7 ошибок. Предлагаемый нами метод LCS дает только 2 ошибки.

Таким образом, достигнутые показатели качества лежат на уровне лучших результатов конкурса по классификации предложений на две группы: парафразы и не парафразы (у лидеров A=0.7459 и F=0.8078 [10]). Наши показатели также вполне сравнимы с результатами, получаемыми на англоязычных текстах. При классификации на три группы наши результаты существенно ниже из-за сложности выделения группы «сомнительных парафраз».

Преимуществом обсуждаемого метода является то, что в процессе работы используется только словарная информация, так что переобучение системы при смене предметной области не требуется.

Литература

[1] Anisimovich, K.V., Druzhkin, K.Ju., Minlos, F.R., Petrova, M.A., Selegey, V.P., Zuev, K.A.: Syntactic and Semantic Parser Based on ABBYY Compreno Linguistic Technologies. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Межд. конф. «Диалог». М.: РГГУ, 2 (11(18)), сс. 91-103 (2012)

[2] Antonova, A.A., Misyurev, A.V.: Russian Dependency Parser SyntAutom at the DIALOGUE-2012 Parser Evaluation Task. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Межд. конф. «Диалог». М.: РГГУ, 2 (11(18)), сс. 104-118 (2012)

[3] Barron-Cedeno, A., Vila, M., Marti, M.A., Rosso, P.: Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. Computational Linguistics, 39 (4), pp. 917-947 (2012)

[4] Corley, C., Mihalcea, R.: Measuring the Semantic Similarity of Texts. Proc. of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, Association for Computational Linguistics Stroudsburg, PA, USA, pp. 13-18 (2005)

[5] Fernando, S., Stevenson, M.: A Semantic Similarity Approach to Paraphrase Detection. Proc. of the Computational Linguistics UK, 11th Annual Research Colloquium (2008)

[6] Finch, A., Hwang, Y.S., Sumita, E.: Using Machine Translation Evaluation Techniques to Determine Sentence-Level Semantic Equivalence. Proc. of the 3rd Int. Workshop on Paraphrasing, pp. 17-24 (2005)

[7] Artemova, G., Boyarsky, K., Gouzйvitch, D., Gusarova, N., Dobrenko, N., Kanevsky, E., Petrova, D.: Text Categorization for Generation of Historical Shipbuilding Ontology. Communications in Computer and Information Science, 468, pp. 1-14 (2014)

[8] http://wordnet.princeton.edu/

[9] http://www.dictum.ru/ru/syntax-analysis/blog

[10] http://www.paraphraser.ru.

[11] http://www.paraphraser.ru/download/get?file_id=5.

[12] Iomdin, L., Petrochenkov, V., Sizov, V., Tsinman, L.: ETAP Parser: State of the Art. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Межд. конф. «Диалог». М.: РГГУ, 2 (11(18)), сс. 119-131 (2012)

[13] Pershina, M., He, Y. Grishman, R.: Idiom Paraphrases: Seventh Heaven vs Cloud Nine. Proc. of the EMNLP 2015 Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pp. 76-82 (2015)

[14] Avdeeva, N., Artemova, G., Boyarsky, K., Gusarova, N., Dobrenko, N., Kanevsky, E.: Subtopic Segmentation of Scientific Texts: Parametr Optimisation. Communications in Computer and Information Science, 518, pp. 3-15 (2015)

[15] Avdeeva, N., Boyarsky, K., Kanevsky, E.: Extraction of Low-frequent Terms from Domain-specific Texts by Cluster Semantic Analyses. Proc. of the ISMW-FRUCT 2016. Saint-Petersburg, Russia, FRUCT Oy, Finland, pp. 86-89 (2016)

[16] Nivre, J., Boguslavsky, I.M., Iomdin, L.L. Parsing the SynTagRus Treebank of Russian. Proc. of the 22nd Int. Conf. on Computational Linguistics, 1, pp. 641-648. Association for Computational Linguistics (2008)

[17] Pham, N., Bernardi, R., Zhang, Y.Z., Baroni, M.: Sentence Paraphrase Detection: When Determiners and Word Order Make the Difference. Proc. of the Towards a Formal Distributional Semantics Workshop at IWCS, pp. 21-29 (2013)

[18] Pronoza, E., Yagunova, E.: Comparison of Sentence Similarity Measures for Russian Paraphrase Identification. Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conf., pp. 74-82 (2015)

[19] Pronoza, E., Yagunova, E., Pronoza, A.: Construction of a Russian Paraphrase Corpus:

Unsupervised Paraphrase Extraction. Proc. of the 9th Russian Summer School in Information Retrieval, August 24–28, 2015, Saint-Petersburg, Russia (RuSSIR 2015, Young Scientist Conference), Springer CCIS

[20] Азарова, И.В., Митрофанова, О.А., Синопальникова, А.А.: Компьютерный тезаурус русского языка типа WordNet. Компьютерная лингвистика и интеллектуальные технологии. Труды Межд. конф. «Диалог 2003». М.: Наука, сс. 43-50 (2003)

[21] Боярский, К.К., Каневский, Е.А.: Предсинтаксический модуль в анализаторе SemSin. Интернет и современное общество: сб. научных статей. Труды XVI Всерос. объединенной конф. «Интернет и современное общество». СПб.: НИУ ИТМО, сс. 280-286 (2013)

[22] Боярский, К.К., Каневский, Е.А.: Семантико-синтаксический парсер SemSin. Научно-технический вестник информационных технологий, механики и оптики, 15 (5), сс. 869-876 (2015)

[23] Боярский, К.К., Каневский, Е.А.: Система продукционных правил для построения синтаксического дерева предложения. Прикладна лінгвістика та лінгвістичні технології: MegaLing-2011. К.: Довіра, сс. 73-80 (2012)

[24] Боярский, К.К., Каневский, Е.А.: Язык правил для построения синтаксического дерева. Интернет и современное общество: Материалы XIV Всерос. объединенной конф. «Интернет и современное общество». СПб.: ООО «МультиПроджектСистемСервис», сс. 233-237 (2011)

[25] Лукашевич, Н.В.: Тезаурус в задачах информационного поиска. М.: МГУ, 495 с. (2011)

[26] Русский семантический словарь. Толковый словарь, систематизированный по классам слов и значений / Российская академия наук. Ин-т рус. яз. им. В.В. Виноградова; Под общей ред. Н.Ю. Шведовой. М.: «Азбуковник» (1998)

[27] Тузов, В.А.: Компьютерная семантика русского языка. СПб: Изд-во С.-Пб. ун-та, 391 с. (2004)