ТЕХНОЛОГИЯ РАЗРАБОТКИ ТЕМАТИЧЕСКИХ СЛОВАРЕЙ НА ОСНОВЕ СОЧЕТАНИЯ ЛИНГВИСТИЧЕСКИХ И СТАТИСТИЧЕСКИХ МЕТОДОВ

Сидорова Е.А.

Российский НИИ Искусственного Интеллекта, г. Новосибирск

lena@iis.nsk.su

Рассматриваемый подход поддерживает технологию автоматического наполнения предметно-ориентированных словарей на основе тематического корпуса текстов, универсального морфологического словаря и набора специальных шаблонов. Словарь включает набор значимых лексем и словокомплексов, их статистические и содержательные характеристики.

Введение

Для существующих методов обработки и анализа текста на естественном языке, как на основе статистических, так и лингвистических подходов [1], особая роль отводится словарю. Если в первом случае словарь содержит информацию о статистических распределениях терминов, то во втором случае он должен включать более содержательную информация, такую как: семантический класс терминов, ассоциативные связи и т.п.

Развитие 2-х подходов идет параллельным курсом, и только в последнее время стали появляться попытки их совместного использования [2]. Классическим примером актуальности данного вопроса служит то, что до последнего времени огромные базы обучающих выборок недостаточно использовались в рамках экспертных (лингвистических) подходов.

Ясно, что требования, предъявляемые к словарю, определяются типом решаемой задачи. С этой точки зрения словари можно подразделить на:

- лингвистические рабочие места [3], которые служат для теоретической работы лингвиста;
- универсальные словари морфологии, современные разработки которых основаны в основном на словаре Зализняка [4, 5];
- статистически-ориентированные словари универсальной лексики, являющиеся основой классических поисковых систем;
- словари, ориентированные на узкую тематику – ПО-словари.

В данной статье рассматривается подход, используемый при разработке технологического комплекса Алекс+, специализированного на создании ПО-словарей. Комплекс позволяет включать в словари как статистическую, так и содержательную информацию и поддерживает технологию автоматического наполнения словаря на основе обучающей выборки.

Архитектура системы

Процесс создания словаря обычно очень трудоемкий процесс, требующий специалистов высокого уровня. В данной работе мы постарались разработать подходы и создать программные средства, облегчающие этот процесс, используя широко известные механизмы статистического сбора информации.

На Рис.1. представлена общая архитектура технологического комплекса Алекс+.

Лексическое наполнение словаря Алекс+ включает словари словоформ, лексем и словокомплексов. Программная оболочка позволяет просматривать и редактировать словари, используя простые средства поиска, сортировки, фильтрации, работы с группой элементов.

Любой термин словаря описывается наборами терминообразующих, статистических и семантических признаков. Сконструировать свой набор признаков для всего словаря или выделенной подгруппы терминов пользователь может с помощью конструктора морфологических типов и редактора иерархии тем.

Однако, если необходимость создания иерархии классов очевидна для разных предметных областей (ПО), то система морфологических классов, как правило, универсальна для выбранного естественного языка. Поэтому в систему изначально введена система морфологических классов, используемая в системе Диалинг [4].

Комплекс включает набор внутренних обработчиков для наполнения и последующей работы со словарем, а также набор внешних обработчиков:

- модуль морфологического анализа Lemmataizer (разрабатываемый в проекте Диалинг);
- модуль внутренней сборки словокомплеков WordFinder;
- модуль настраиваемой сборки сложных структур, на основе системы правил-шаблонов Alex-T (также разрабатываемой нашим институтом).

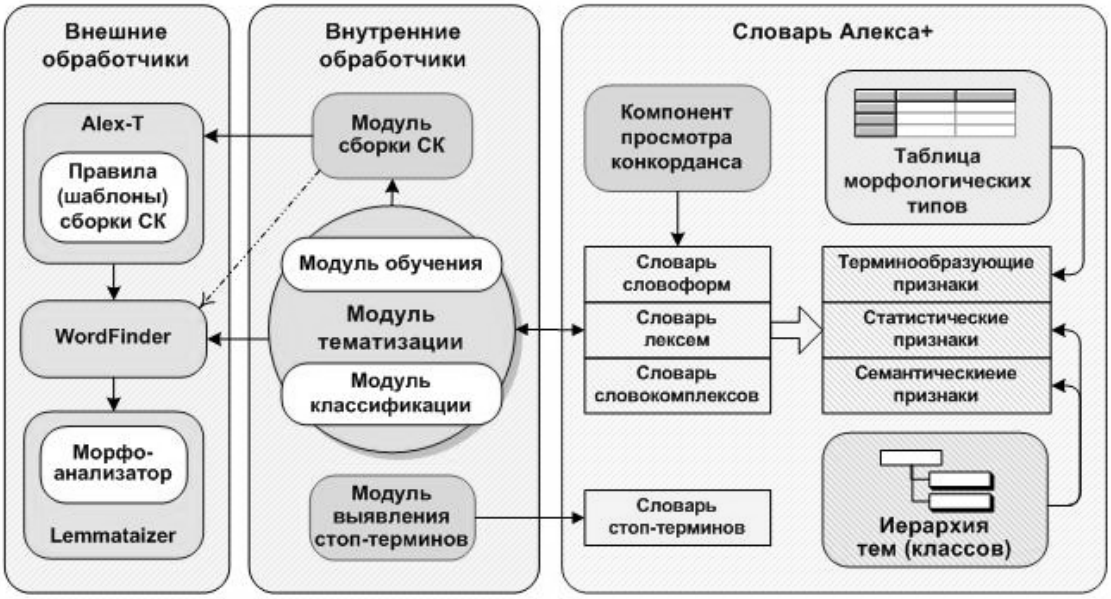

Рис. 1. Архитектура системы Алекс+

Модуль тематизации обеспечивает анализ текста в различных режимах: наполнение словаря, ведение статистики встречаемости терминов, классификация текста на основе статистики. Последовательный анализ текста в разных режимах позволяет поддерживать механизм расширения иерархии классов и «дообучения» словаря.

Представление словарной статьи

Рассматриваемую технологию можно разделить на два взаимосвязанных, но качественно различных режима работы:

- конструкторский: режим создания и отладки словаря,
- рабочий: режим применения созданного и отлаженного словаря для решения соответствующих задач обработки текста, например для классификации или индексации.

В конструкторском режиме словарь включает: словарь словоформ, словарь лексем, словарь словокомплексов (СК) и словарь стоп-терминов. В рабочем режиме словарь не содержит стоп-терминов и не может быть изменен (но он может быть использован для создания других словарей в конструкторском режиме). Это требование необходимо для оптимизации скорости и ресурсов при работе со словарем.

Любой термин словаря содержит: терминообразующие, статистические и семантические признаки, а также статус, предоставляющий пользователю возможность следить за обновлением словаря. Статус позволяет выявлять новые, обновленные (если статистика была изменена), измененные и введенные вручную понятия.

а. Терминообразующие признаки

Терминообразующие признаки (ТП) служат для того, чтобы с одной стороны выявить понятие в тексте (анализ), с другой – послужить основой для построения терминов (синтез).

Для разных типов терминов набор ТП различен. Так, лексема содержит следующие признаки:

- нормальная форма,
- морфологический класс (класс, в частности, включает часть речи и словообразующие морфологические признаки лексемы),
- основа – неизменяемая часть лексемы (у некоторых лексем основа может быть пустой, тогда лексема определяется парадигмой),
- парадигма – набор псевдофлексий слова,
- тип (правило) образования: слово универсального словаря, предсказание, слово служебного словаря, слово из словаря имен и т.п.

Ясно, что основа в сочетании с парадигмой лексемы дают все ее словоформы. Каждой словоформе в словаре сопоставлен набор словоизменяющих морфологических признаков, определяемых морфологическим классом лексемы.

СК содержит следующие признаки:

- нормальная форма СК,
- список составляющих терминов (в состав СК может входить другой СК),
- правило (имя шаблона), согласно которому образуется данный СК.

б. Статистические признаки

Статистические признаки накапливают статистическую информацию о появлении термина в обрабатываемых текстах. Для ведения статистики необходимо, чтобы пользователь задал систему связанных между собой тем (система тем может

быть пустой, тогда статистика будет вестись только по одной теме, отражающей всю выборку), а также наличие обучающего корпуса текстов.

В словаре для каждого термина хранится:

1) встречаемость в обучающей выборке,
2) количество текстов выборки, в которых хотя бы 1 раз встретилось данной понятие,
3) список тем, в которых встретился термин (без учета наследования),
4) встречаемость и количество текстов выборки для каждой темы.

Остальные статистические параметры (частота встречаемости в выборке, а также частота и вес по каждой теме) вычисляются динамически:

- частота употребления каждого термина вычисляется как отношение количества словоупотреблений данного термина в выборке к количеству всех словоупотреблений в выборке;
- частота употребления термина в теме (показатель, насколько данный термин выражает тему в сравнении/относительно других слов темы), вычисляется как отношение количества словоупотреблений данного термина в теме к количеству всех словоупотреблений в теме;
- вес термина в теме (показатель, как часто данный термин соотносится теме), отношение частоты термина во всей выборке к частоте термина в теме.

Как частота, так и вес термина по теме нормируются по длине выборки.

в. Семантические признаки

Под семантическим признаком понимается произвольный атрибут простого типа, задаваемый пользователем вручную для всех или выделенной группы терминов словаря. Кроме того, к семантическому признаку относится принадлежность к некоторому классу или классам.

На данный момент иерархия классов совпадает с иерархией тем и предложен механизм семантического соотнесения любого термина теме (классу) осуществляемого пользователем вручную.

В дальнейшем планируется реализация механизма загрузки иерархии классов из некоторого внешнего источника (онтологии), которая будет существовать наравне с иерархией тем.

Механизмы настройки структуры словаря

Одной из задач разрабатываемой технологии было создание гибких механизмов позволяющих специалисту проводить тонкую настройку структуры словаря. В данной главе рассматриваются способы создания и редактирования морфологических классов, морфологических атрибутов, типов парадигм, а также принципы создания иерархии тем.

а. Морфологические типы

Описание морфологической информации включает описания следующих понятий:

- морфологический атрибут,
- часть речи,
- тип парадигмы,
- морфологический тип (класс).

Морфологический атрибут описывается именем и множеством своих значений $<N_i, X_i>$ (например, <Род, {муж, жен, ср}>). Часть речи тоже является атрибутом, но так как она всегда должна присутствовать при описании типа, то было принято решение вынести ее в отдельную сущность.

Тип парадигмы задает длину парадигмы данного типа и сопоставляет каждому элементу парадигмы набор значений атрибутов (например, для «простого» прилагательного это тройка <падеж, число, род>). Такой набор строго упорядочен, что позволяет использовать компактную форму записи в виде древообразной структуры, вершинами которой являются подмножества множеств значений атрибутов $<A_i, X_i>$. Пара функций $f:n \rightarrow X_{i1}*\ldots*X_{ik}$, $g:X_{i1}*\ldots*X_{ik} \rightarrow n$ по любой заданной структуре обеспечат преобразование индекса или номера псевдофлексии в парадигме в набор значений атрибутов, и наоборот.

Т.о. каждой лексеме сопоставляется некоторая парадигма из таблицы парадигм, а каждой парадигме сопоставляется тип парадигмы, описывающий ее структуру.

Морфологический тип описывает морфологический класс лексемы и, вместе с основой, однозначно идентифицирует лексему. Тип включает часть речи, набор признаков лексемы ($x_{ij} \epsilon X_i$) и тип парадигмы, описывающий признаки словоформ лексемы.

б. Иерархия тем

Система тем (классов) задается в виде набора классов и связывающих их отношений наследственности (включая множественное наследование). Иерархия имеет следующие конструктивные ограничения:

- отсутствие циклов, т.к. отношение наследственности транзитивно, не должно быть ситуации, когда тема является сама себе родителем;
- для любой темы в списке непосредственных родителей не должно быть двух тем, которые связаны между собой отношением наследственности. Наличие такой ситуации повлекло бы за собой семантическую непрозрачность системы описания мира.

Для каждой темы хранится следующая статистическая информация:

1) количество текстов данной темы в обучающей выборке,
2) количество терминов (лексем и СК) во всех текстах данной темы.

Статистика по темам накапливается без учета наследования (учитывать наследование, в зависимости от задачи, можно будет динамически).

Модули автоматизированной настройки словаря

Предлагаемая технология ориентирована на две основные группы пользователей: лингвистов и экспертов в заданной ПО, и содержит набор модулей, позволяющих автоматизировать процесс наполнения и настройки словаря.

а. Сборка словокомплексов

Для первоначальной сборки СК используется модуль WordFinder, который осуществляет автоматическую сборку СК по фиксированному набору правил. В качестве морфоанализатора WordFinder использует внешний морфологический модуль Lemmatazer системы Диалинг.

К недостаткам модуля можно отнести нерасширяемость системы правил, а также неизменяемость структуры таблицы морфологических классов, полученной от Lemmataizer-а (как уже было сказано, система Алекс+ позволяет изменять существующие морфологические классы и все изменения должны отражаться в словаре, сопоставляющем морфологические классы системы Lemmataizer с внутренними классами).

Для возможности расширения системы правил был подключен еще один внешний модуль – Алекс-Т, который позволяет пользователю описывать правила формирования СК как для классов лексем, так и для отдельно взятого набора лексем с помощью механизма шаблонов. На вход модулю поступает цепочка объектов, сформированная либо модулем WordFinder, либо непосредственно словарем Алекса+, со всеми морфологическими признаками (заданными уже в терминах признаков Алекса+). Результатом работы Алекса является набор объектов, построенных на основании шаблонов, со следующей информацией:

- имя правила сборки СК;
- набор составных объектов;
- информация для синтеза уникального имени СК (синтез имени происходит непосредственно в словаре Алекса+ на основании всей морфологической информации о лексемах, входящих в состав СК).

б. Общая схема обучения

Под обучением понимается процесс формирования словаря со статистическими показателями, т.е. словаря, элементам которого сопоставляется статистическое распределение по классам (темам). Обучение происходит на основе обучающей выборки – массива текстов с исходной разметкой принадлежности к темам.

Можно выделить следующие этапы обучения:

1) Морфологический анализ текста и сборка СК. Результатом работы этого этапа является выделение списка значимых лексем и СК.
2) Каждый термин, обнаруженный в тексте темы t ищется в словаре и, если он там не находится, то добавляется. В результате этой операции корректируются статистические показатели термина и темы t.
3) В результате обработки всех текстов обучающей выборки для «значимого» словаря строится матрица, столбцы которой соответствуют классам, а строки – лексемам и СК. Ячейки этой матрицы на пересечении термина x и класса t будут отражать коэффициент вероятности отнесения текста, включающего термин x, к классу t.
4) Теперь для каждого текста выборки можно при помощи модуля классификации определить, к какой (каким) теме он относится.
5) Полученное достаточно грубое распределение весов анализируется лингвистом для уточнения и исключения случаев, ухудшающих распознавание. При этом используется конкорданс и выборка случаев, получивших неправильную или недостаточно ясную классификацию.

в. Модуль классификации

Наличие словарных статистических показателей делает возможным применение классических методов классификации – процесса распознавания темы (набора тем) текста.

На данном этапе развития системы используется простейший случай, а именно: для всех значимых терминов из таблицы по каждому классу (теме) берется суммы весов тех терминов, веса которых превышают шумовой уровень лексики, и вычитается сумма обратных весов тех терминов, веса которых ниже шумовой уровень лексики. Т.о. при анализе учитывается не только «положительная», но и «отрицательная» информация о соответствии термина теме. Текст относится к теме\темам, получившим значение функции выше некоторого порога.

С развитием этого этапа можно будет от простых функций распознавания, переходить более сложным, учитывающих корреляцию пар лексем, наличие словокомплексов, выявление сложной значимой лексики и конструкций.

г. Тематизация

Главным недостатком автоматического обучения является то, что пользователь сразу должен задать иерархию тем, по которой размечается обучающая выборка. Однако на практике, типичной является ситуация, когда требуется расширять и углублять существующую иерархию.

Поочередно используя механизмы классификации и «дообучения» можно дать пользователю возможность расширять иерархию

тем. Средства, реализующие этот механизм, получили название модуль тематизации.

Частным случаем такой ситуации является создание словаря, иерархии тем и обучающей выборки «с нуля». Рассмотрим основные этапы этого механизма:

1) Анализ множества неразмеченных тестов и лексическое наполнение словаря.
2) Пользователь вручную просматривает словарь (используя механизмы фильтрации по статусу, встречаемости и т.п.) и выделяет набор ключевых терминов «маркирующих» новые темы.
3) Автоматическая «грубая» классификация неразмеченных текстов и анализ пользователем полученных результатов с целью доуточнения разметки текстов.
4) Последним шагом является обучение оставшейся части словаря.

Если результат шага 3 совершенно не устраивает пользователя, он может еще раз перейти ко 2-му этапу, дополнить словарь и заново осуществить разметку текстов, либо вернуться ко 2-му шагу после 4-го. Т.о. пользователь может использовать этот механизм до тех пор, пока результат его не удовлетворит.

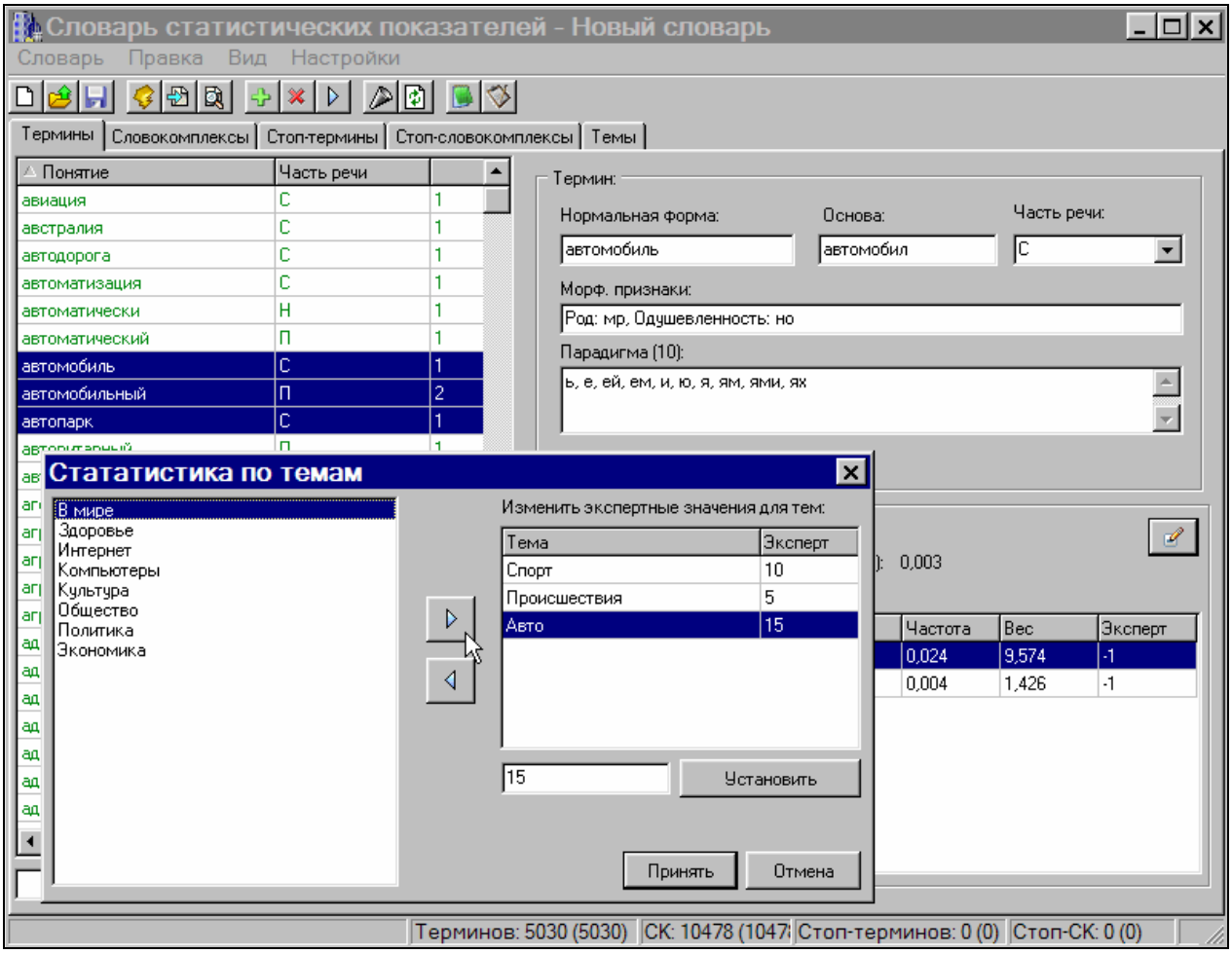

Рис. 2. Словарь – интерфейс пользователя

Рис. 2. иллюстрирует шаг 2, где пользователь вручную может отмечать ключевые термины.

Отметим, что механизм тематизации обеспечивается модулями обучения и классификации, но должен самостоятельно отслеживать изменения статистики (чтобы не было ее дублирования при повторных обработках текстов), а также управлять режимами обучения (с/без накопления статистики, с/без добавления новых терминов) с целью оптимизации работы модуля.

Выявление стоп-терминов

Словарь стоп-терминов (стоп-словарь) содержит с одной стороны шумовую общеупотребительную лексику, с другой стороны – исключения или ошибочные термины, возникновение которых связано с одной из следующих причин:

- неправильное предсказание лексемы по встретившемуся в тексте незнакомому слову;

- неправильно собранный СК, по причине неучтенных тонкостей языкового выражения, заданных с помощью шаблонов на классах лексем;
- грамматические ошибки написания, встретившиеся в обучаемой выборке.

Механизм автоматического определения шумовых стоп-терминов основан на статистическом распределении веса термина по рубрикам (темам). Чтобы термин можно было отнести к стоп-словарю его вес должен лежать в некотором пороговом интервале для всех тем (по которым накоплена достаточная статистическая информация).

Выявление ошибок обычно происходит вручную. Для облегчения ручной обработки эксперту предоставляется возможность фильтровать и просматривать часть словаря по любым статистическим параметрам. Так, при большом объеме обучающей выборки, ошибки третьего типа выявляются при фильтрации словаря по параметру «встречаемость в выборке» в интервале от 1 до 5, а при обнаружении неправильного словокомплекса, стоит посмотреть на все словокомплексы собранные по тому же правилу (что в идеальном случае должно привести к редактированию самого правила).

Словарь стоп-терминов требуется на стадии обучения, или, если предполагается, что словарь в дальнейшем будет пополняться не вручную. В этом случае, при очередной обработке текста предварительно будет проверяться наличие термина в стоп-словаре, прежде чем этот термин будет как-либо использован.

Заключение

Словарь, созданный с помощью предложенной технологии, может поддерживать основные этапы анализа текста: морфологический, синтаксический и семантический, а также классификацию на основе статистики.

В данный момент реализован прототип системы, который планируется использовать для задач индексации и классификации порталов по археологии и искусственному интеллекту, а также для развития системы интеллектуализации документооборота InDoc[6].

Расширение и доработка системы будет заключаться в оптимизации скорости обработки текстов, увеличение порога ограничения на объем словаря, реализация полновесного банка текстов, реализация более сложных функций классификации и улучшения качества анализа текста с применением поверхностного семантического анализа. Следующим этапом развития подхода является разработка и создание конструктора онтологий, который бы позволил объединить все настраиваемые элементы базы знаний в одном пользовательском интерфейсе.

Список литературы:

1) Хорошевский В.Ф. Управление знаниями и обработка ЕЯ-текстов // Труды 9-й национальной конференции по искусственному интеллекту КИИ’2004. М.: Физматлит, 2004. Т.2, С.565–572.
2) Агеев М.С., Добров Б.В., Лукашевич Н.В. Поддержка системы автоматического рубрицирования для сложных задач классификации текстов // RCDL’2004 Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Пущино, 2004.
3) Сулейманов Д.Ш., Гатиатулин А.Р. Структурно-функциональная компьютерная модель татарских морфем // Казань: ФЭН, 2003.
4) Сокирко А.В. Морфологические модули на сайте www.aot.ru // Труды международного семинара Диалог’2004 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2004.
5) http://www.rco.ru/ – RCO (Russian Context Optimizer). Технологии анализа и поиска текстовой информации.
6) Кононенко И.С., Сидорова Е.А. Обработка делового письма в системе документооборота // Труды международного семинара Диалог’2002 по компьютерной лингвистике и ее приложениям. М.: Наука, 2002. Т.2, С. 299–310.