OSTIS-2015
(Open Semantic Technologies for Intelligent Systems)

УДК 004.892

ОЦЕНКА ТЕРМИНОЛОГИЧНОСТИ ЛЕКСИЧЕСКИХ ЕДИНИЦ НА ОСНОВЕ ОНТОЛОГИИ ПРЕДМЕТНОЙ ОБЛАСТИ

Андреев И.А., Башаев В.А., Клейн В.В., Мошкин В.С., Ярушкина Н.Г.

Ульяновский государственный технический университет,
г. Ульяновск, Российская Федерация

ares-ilya@yandex.ru
perevod73@yandex.ru
vikklein93@gmail.com
postforvadim@yandex.ru
jng@ulstu.ru

В данной статье описана семантическая метрика извлечения списка терминов из текстов конкретной проблемной области, основанная на анализе ее онтологии. Представлена формальная модель используемой OWL-онтологии, также рассмотрена реализация моделей и алгоритмов оценки степени терминологичности слов или сочетаний слов текстовых массивов в программной системе извлечения терминологии из текста.
Ключевые слова: извлечение терминов, семантическая метрика, онтология.

Введение

Принцип работы существующих алгоритмов извлечения терминологии (term extraction) в лексикографии и терминоведении основан на статистических и лингвистических методах. В основе статистических методов лежит вычисление степени терминологичности на основании числовых закономерностей, присущих термину или нетермину. В основе лингвистических методов лежит отбор по определенным лексико-грамматическим шаблонам и другим лингвистическим признакам термина [Андреев и др., 2013].

Главным недостатком использования статистических и лингвистических методов в процессе извлечения терминологии из текста является отсутствие возможности выделения из получившегося множества терминов только тех, которые относятся к рассматриваемой проблемной области [Yarushkina, 2001].

На множестве информационных единиц в некоторых случаях полезно задавать отношение, характеризующее ситуационную близость информационных единиц, т. е. силу ассоциативной связи между информационными единицами. Его можно было бы назвать отношением релевантности для информационных единиц [Namestnikov et al., 2002].

При анализе больших массивов документации необходимо учитывать специфику ее предметной области, чтобы получить в качестве результата список терминов, характерных для конкретной предметной области. Именно для решения подобных задач используют семантические алгоритмы, базирующиеся на определенных семантических метриках.

В настоящее время одной из наиболее универсальных методик представления экспертных знаний с точки зрения полноты семантического описания информационной единицы предметной области является онтологический подход. Именно поэтому одним из важнейших направлений решения задачи извлечения терминологии из большого массива технической документации является разработка и использование семантических метрик на основе онтологических моделей [Ярушкина и др., 2007а].

1. Формальная модель онтологии предметной области "эксплуатация токарно-фрезерного станка с ЧПУ"

Сущность онтологического подхода заключается в том, что предметная область представляется в виде организованной совокупности понятий, их свойств и связей.

Наиболее удобным форматом представления онтологии с точки зрения машинной обработки и

наглядности описания особенностей предметной области является язык OWL.

Выделим обязательные требования к OWL-онтологии, используемой в рамках решения задачи извлечения терминологии:

- Онтология должна наиболее полно отражать особенности объектов предметной области.
- Онтология не должна быть избыточной.
- Онтология должна быть наглядной.

Онтологический подход хранения знаний предполагает представление их в следующем виде:

$$O = < T, R, F> \quad (1)$$

Исходя из модели (1), онтология «Эксплуатация токарно-фрезерного станка с ЧПУ» имеет следующие составляющие:

1. T – термины прикладной области, которую описывает онтология. Например, объекты «Резцедержатель», «Станина», «Поплавковое реле».
2. R – отношения между терминами предметной области, при этом $R \subset \{R_{inc}, R_{add}, R_{term}, R_{lem}, R_{NC}\}$:

- R_{inc} – множество встроенных отношений объектов таких, как «sameAs» и «SubClassOf»;
- R_{add} – множество отношений, позволяющих расширять набор объектов описываемой предметной области за счет сочетания лемм связанных объектов;
- R_{term} – отношение «является Термином», имеющее логический тип значения. Это свойство является вспомогательным и определяется экспертом исходя из критерия, насколько данный объект онтологии является характерным конкретно для этой предметной области;
- R_{lem} – отношение «имеет Лемму», имеющее строковое значение, полученное путем леммирования (приведения к начальной форме) наименования объекта с помощью программы Mystem компании Яндекс по соответствующим морфологическим признакам термина;
- R_{NC} – множество отношений объектов, а также свойств типа данных, наиболее полно описывающих особенности взаимодействия объектов рассматриваемой предметной области.

3. F – множество функций интерпретации (аксиоматизации), заданных на терминах и/или отношениях онтологии [Добров и др., 2003]. Примеры таких функций в разработанной онтологии представлены выражениями (2) и (3):

$$F_{СОЖ} : X_{ТипСверления} \rightarrow Y_{ТипПодачи} \quad (2)$$

где $F_{СОЖ}$ – отношение «является Типом Подачи СОЖ»,

$X_{ТипСверления}$ – множество объектов класса «Тип Сверления»,

$Y_{ТипПодачи}$ – множество объектов класса «Тип Подачи СОЖ».

$$F_{InEng} : X_{\text{Контекст}} \rightarrow Y_{Eng} \quad (3)$$

где F_{InEng} – отношение «имеет Английский Эквивалент»,

$X_{Контекст}$ – множество объектов класса «Контекст»,

Y_{Eng} – множество объектов класса «Английский Аналог».

2. Семантическая метрика оценки терминологичности слов/сочетаний слов

Использование семантической метрики «термин/нетермин» на множестве слов конкретного текста с использованием заранее разработанной OWL-онтологии в процессе извлечения терминологии предполагает определение для каждого поступающего слова или сочетания слов степени близости к терминам рассматриваемой области.

Степень близости входных слов/сочетаний слов к терминам проблемной области (k_{Ont}) может иметь значение от 0 до 1: чем ближе полученное значение к 1, тем с большей долей вероятности данное одно-/многословие является термином [Афанасьева и др., 2011].

В ходе решения поставленной задачи было разработано два критерия выделения терминов из предметной области посредством использования онтологии:

- тезаурусный критерий,
- критерий вложенных связей.

Результаты проведенных экспериментов должны показать, какой из данных семантических критериев является наиболее продуктивным и оптимальным применимо к имеющейся модели онтологии.

2.1. Тезаурусный критерий

Тезаурус представляет собой контролируемый словарь терминов на естественном языке, явно указывающий отношение между терминами и предназначенный для информационного поиска. Любая онтология является усложненной версией тезауруса [Кураленок и др., 2002].

Тезаурусный подход к извлечению терминологии предполагает непосредственный поиск вхождений лемм поступающих на вход слов и их сочетаний среди терминов, определенных в онтологии. Для этого в разработанной онтологии для каждого класса определено свойство «имеет Лемму», которое имеет строковое значение, полученное путем леммирования (приведения к начальной форме) имени объекта с помощью программы Mystem компании Яндекс по соответствующим морфологическим признакам термина.

Алгоритм определения степени близости слов/сочетания слов терминам проблемной области согласно тезаурусному критерию предполагает:

- Оценку степени близости поступающего на вход алгоритма слова/сочетания слов каждому объекту онтологии без учета онтологического критерия оценки;
- Определение опорного объекта онтологии, наиболее близко ассоциирующегося с входным одно-/многословием.

Опорный объект онтологии, используемый в дальнейшем анализе, имеет степень близости по отношению к входному слову/сочетанию слов, рассчитанную по следующей формуле:

$$k_t = \max_{i=1}^{m}(\frac{n_i}{p_i}) \qquad (4)$$

где m – количество всех объектов онтологии;

n_i – число слов из леммы входного многословия, найденных в лемме объекта онтологии;

p_i – общее число слов в лемме объекта онтологии.

Общая схема оценки степени близости слов/сочетания слов терминам проблемной области согласно тезаурусному критерию приведена на рисунке 1.

Рисунок 1 – Поиск опорного объекта онтологии

При этом порядок следования слов многословия в опорном объекте должен сохраняться.

Если несколько разных объектов онтологии имеют одинаковое значение коэффициента k_t, то опорным будет считаться тот объект, которому соответствует максимальное n_i. Если таких объектов несколько, то они все будут считаться опорными и анализ по онтологическому критерию будет проведен для каждого из этих объектов.

Структура онтологии предполагает наличие у каждого из ее объектов свойства (DatatypeProperty) «является Термином», имеющее логический тип значения. Это свойство является вспомогательным и определяется экспертом исходя из критерия, насколько данный объект онтологии является характерным конкретно для этой предметной области.

Степень близости слова/сочетания слов терминам рассматриваемой предметной области в соответствии с тезаурусным критерием оценивается по следующей формуле:

$$k_{Ont} = \frac{k_t}{c+1} \qquad (5)$$

где k_t – результат первого этапа анализа; c – число отношений, связывающих опорный объект онтологии с ближайшим объектом, имеющим истинное значение свойства «является Термином».

В случае, если сам опорный объект имеет истинное значение данного свойства, то $c = 0$.

Таким образом, процесс оценки степени близости одно/многословия к терминам проблемной области по метрике «термин/нетермин» в его онтологической составляющей представляет собой движение по графу, в узлах которого находятся объекты соответствующих классов онтологий.

2.2. Критерий вложенных связей

Помимо оценки степени терминологичности отдельно взятого слова/сочетания слов, разработанная метрика позволяет извлечь термины из текста посредством их сопоставления с имеющимися объектами и сочетаниями лемм соответствующих объектов с помощью отношений R_{add}, определенных в онтологии.

Таким образом, при сопоставлении входных сочетаний и объектов предметной области, связанных между собой однонаправленными отношениями R_{add}, термином рассматриваемой предметной области будет считаться многословие, лемма которого полностью совпадает с объединением лемм соответствующих объектов онтологии.

Определяющими для использования этого метода являются отношения R_{add}, связь объектов посредством которых позволяет формировать словосочетания естественным образом. Пример формирования многословий с помощью свойства «имеет Отношение»:

1. Найденная цепочка объектов: «Вращение» + «имеет Отношение» + «Двигатель» + «имеет Отношение» + «Переменный ток»;
2. Объединение лемм объектов онтологии: «вращение двигатель переменный ток»;
3. Термин, извлекаемый из обрабатываемого текста: «вращение двигателя переменного тока».

При этом извлеченные термины, входящие в свою очередь в термины, состоящие из большего количества слов, не рассматриваются в качестве терминов с целью избегания избыточности.

2.3. Метрики оценки результатов

Рассмотрим метрики оценки, применимые к задаче классификации. Сгруппируем ответы нашего гипотетического анализатора следующим образом:

- Истинно-положительные (**true positives, tp**) – те категории, которые мы ожидали увидеть и получили на выходе;
- Ложно-положительные (**false positives, fp**) – категории, которых быть на выходе не должно, и анализатор их ошибочно вернул на выходе;
- Ложно-отрицательные (**false negatives, fn**) – категории, которые мы ожидали увидеть, но анализатор их не определил;
- Истинно-отрицательные (**true negatives, tn**) – категории, которых быть на выходе не должно, и на выходе анализатора они тоже совершенно правильно отсутствуют.

В этом случае мера точности (*P, precision*) определяется так:

$$P = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fp}} \qquad (6)$$

Мера точности характеризует, сколько полученных от классификатора положительных ответов являются правильными.

Мера точности, однако, не дает представление о том, все ли правильные ответы вернул классификатор. Для этого существует так называемая мера полноты (*R, recall*):

$$R = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fn}} \qquad (7)$$

Мера полноты характеризует способность классификатора «угадывать» как можно большее число положительных ответов из ожидаемых [Ярушкина и др., 2014].

Помимо этого, удобно для характеристики классификатора, использующего разработанную семантическую метрику, использовать унифицированную метрику F_1:

$$F_1 = 2 \times \frac{P \cdot R}{P + R} \qquad (8)$$

Фактически это просто среднее гармоническое величин P и R. Именно F_1 используется, чтобы сформулировать пороговое качество разработанного семантического анализатора [Ярушкина и др., 2007b].

3. Семантическая метрика оценки терминологичности слов/сочетаний слов

В рамках решаемой поставленной задачи были проведены следующие действия:

1. Экспертом в области эксплуатации токарно-фрезерного станка с числовым программным управлением построена OWL-онтология соответствующей проблемной области;
2. Была разработана онтологически-ориентированная система извлечения терминологии, применяющая описанные выше метрики для решения задачи определения терминологичности одно/многословий, извлекаемых из больших объемов технических текстов.

3.1. Онтологически ориентированная система извлечения терминологии

Разработанная OWL-онтология имеет иерархическую организацию и включает в себя 261 экземпляр классов и порядка 746 отношений объектов классов. На данный момент онтология имеет 4 уровня иерархии, что позволяет максимально конкретизировать термины предметной области, используемой при решении поставленной задачи.

Пример объявления класса «Концевая фреза» разработанной OWL-онтологии:

```
<owl:Class rdf:ID="Концевая_фреза ">
  <rdfs:label
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >Концевая фреза</rdfs:label>
  <rdfs:subClassOf>
   <owl:Class rdf:about="#Фреза"/>
  </rdfs:subClassOf>
</owl:Class>
```

Для реализации описанных алгоритмов была разработана программная система «Онтологически-ориентированная система извлечения терминологии».

Алгоритм работы разработанной системы извлечения терминологии предполагает следующую последовательность действий:

1. Обработка текста модулем морфологического анализа;
2. Подсчет лингвистических и статистических характеристик полученного текста, содержащего морфологическую разметку, основным модулем системы;
3. Подсчет семантических характеристик слов и словосочетаний обрабатываемого текста, базирующийся на представленных методиках с использованием разработанной OWL-онтологии.

3.2. Результаты вычислительных экспериментов извлечения терминов из текстов соответствующей предметной области

Приводимые результаты экспериментов имеют целью изучение эффективности разработанных показателей. Были рассмотрены результаты работы двух показателей: тезаурусного и критерия внутренних связей; четырех категорий словоупотреблений: одиночных слов, двух-, трех-, четырехсловных словосочетаний.

Для проведения эксперимента использовался текст объемом около 62000 слов из руководства по эксплуатации токарно-фрезерного станка с ЧПУ.

К особенностям текстов данной предметной области можно отнести высокую насыщенность терминами, влияние научного стиля на лексико-семантические, морфологические, синтаксические параметры и формализованность содержания, опирающегося на логико-понятийную схему предметной области.

Для оценки эффективности подсчета показателей рассмотрены меры *Precision* (6), *Recall* (7) и F_1-мера (8) для каждого показателя в каждой категории словоупотреблений.

Результаты экспериментов по извлечению терминов посредством применения тезаурусного критерия представлены в таблице 1, критерия вложенных связей – в таблице 2.

Так как в случае применения тезаурусного критерия оценивается терминологичность каждого слова/сочетания, поступающего на вход алгоритма, то для формального отделения терминов от нетерминов в результате его выполнения, необходимо ввести пороговое значение $k_{Ont} = 0{,}5$.

Таблица 1 – Результаты применения тезаурусного критерия

Кол-во слов	Термины	k_{Ont} >0,5	Верно	*P*	*R*	F_1-мера
1	294	120	88	0,73	0,29	0,42
2	631	305	133	0,43	0,21	0,28
3	361	379	214	0,56	059	0,57
4	107	196	120	0,61	1,12	0,79

Таблица 2 – Результаты применения критерия вложенных связей

Кол-во слов	Термины	Кандидаты	Верно	*P*	*R*	F_1-мера
1	294	168	154	0,91	0,52	0,66
2	631	431	372	0,86	0,58	0,69
3	361	370	327	0,88	09	0,89
4	107	159	129	0,81	1,2	0,97

Анализ результатов выполнения разработанных методик необходимо рассматривать с учетом различий вариантов словоупотреблений:

3.2.1. *Одиночные слова*

Исходя из полученных выше результатов, следует отметить, что наилучшие показатели извлечения однословных терминов были получены при применении второго критерия. Почти все извлеченные алгоритмом одиночные слова являются терминами, в то время как всего было извлечено немногим более половины всех однословных терминов рассматриваемой проблемной области. *Recall* у тезаурусного показателя для однословных терминов хоть и ненамного ниже, но значение *Recall* позволяет судить о более низкой эффективности этого показателя. Таким образом, показатель вложенных связей оказался наиболее эффективным при вычислении однословных терминов, о чем свидетельствует и наивысшее значение F_1-меры среди показателей.

3.2.2. *Двухсловные словосочетания*

Исходя из результатов анализа, можно сделать вывод, что тезаурусный признак значительно уступает по полноте и точности второму критерию, имеющему лучшие значения *Precision*, *Recall* и F_1-меры среди всех результатов. При достаточно высокой точности критерий вложенных связей извлекает более половины двухсловных терминов предметной области. Таким образом, для извлечения двухсловных терминов наиболее эффективным также является второй критерий.

3.2.3. *Трехсловные словосочетания*

Удовлетворительным можно считать результат работы тезаурусного признака по извлечению трехсловных терминов: извлекается более половины трехсловных терминов предметной области при среднем *Precision*. Результаты работы второго критерия можно назвать лучшими, о чем позволяют судить достаточно высокие значения *Precision* и *Recall*.

3.2.4. *Четырехсловные словосочетания*

Тезаурусный признак и критерий вложенных связей оказались сопоставимыми по эффективности. Значение *Recall,* превышающее 1 для обоих показателей, свидетельствует об извлечении ими терминов, ранее не выделенных в ходе экспертного анализа. Несмотря на схожие результаты, тезаурусный признак проигрывает второму признаку за счет более низкого значения *Precision*. Таким образом, признак вложенных связей оказался наиболее эффективным для извлечения и четырехсловных терминов.

Полученные результаты экспериментов по извлечению терминов из инструкции по эксплуатации токарно-фрезерного станка с ЧПУ с использованием разработанной онтологии соответствующей предметной области позволяют сделать вывод о высокой эффективности использования критерия вложенных связей для решения поставленной задачи, особенно в случаях анализа трех- и четырехсловий.

Заключение

Таким образом, предложенная в данной работе семантическая метрика «термин/нетермин» на основе онтологии проблемной области позволяет выделить из массива поступающих одно-/многословий только те термины и сочетания, которые относятся к данной предметной области, устанавливая для каждого из входных сочетаний слов численное значение степени их близости к терминам рассматриваемой предметной области.

Данная метрика может быть использована как в качестве самостоятельной, так и в сочетании с лингвистическими и статистическими метриками, используемыми в процессе извлечения терминологии с целью обеспечения всестороннего анализа поступающих данных.

Библиографический список

[Андреев и др., 2013] Андреев И.А., Башаев В.А., Клейн В.В. Разработка программного средства для извлечения терминологии из текста на основании морфологических признаков, определяемых программой Mystem // Интегрированные модели и мягкие вычисления в искусственном интеллекте. – М. : Физматлит, 2013. – С. 1227–1236.

[Yarushkina, 2001] Yarushkina N. Soft computing and complex system analysis // International Journal of General Systems. – 2001. – Vol. 30, № 1. – pp. 71–88.

[Namestnikov et al., 2002] Namestnikov A., Yarushkina N. Efficiency of Genetic algorithms for automated design problems // Известия Российской академии наук. Теория и системы управления. – 2002. – № 2. – С. 127–133.

[Ярушкина и др., 2007a] Ярушкина Н.Г., Вельмисов А.П., Стецко А.А. Средства data minig для нечетких реляционных серверов данных // Информационные технологии. – 2007. – № 6. – С. 20–29.

[Добров и др., 2003] Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических словосочетаний по текстам предметной области // Тр. 5-й Всеросс. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL-2003). – СПб., 2003. – С. 201–210.

[Афанасьева и др., 2011] Афанасьева Т.В., Ярушкина Н.Г. Нечеткий динамический процесс с нечеткими тенденциями в анализе временных рядов // Вестник Ростовского государственного университета путей сообщения. – 2011. – Т. 3. – С. 7–16.

[Кураленок и др., 2002] Кураленок И.Е., Некрестьянов И.С. Оценка систем текстового поиска // Программирование. – 2002. – 28(4). – С. 226–242.

[Ярушкина и др., 2014] Ярушкина Н.Г., Мошкин В.С Применение онтологического подхода к анализу состояния локальной вычислительной сети // Радиотехника. – 2014. – № 7. – С. 120–124.

[Ярушкина и др., 2007b] Ярушкина Н.Г., Афанасьева Т.В. Нечеткие временные ряды как инструмент для оценки и измерения динамики процессов // Датчики и системы. – 2007. – № 12. – С. 46–50.

ALGORITHMS FOR EVALUATION OF WORD COMBINATIONS OR WORDS MEMBERSHIP DEGREE TO TERM LIST BASED ON SUBJECT AREA ONTOLOGY

I.A. Andreev., V.A. Bashaev, V.V. Klein, V.S. Moshkin, N.G. Yarushkina.

Ulyanovsk State Technical Uuniversity, Russia

ares-ilya@yandex.ru

perevod73@yandex.ru

vikklein93@gmail.com

postforvadim@yandex.ru

jng@ulstu.ru

This article describes a semantic metric retrieve a list of terms from the texts for this specific problem, based on an analysis of its ontology. A formal model used OWL-ontologies, as well as models and algorithms for assessing membership degree of word or combinations of words to term list.

In addition, the metrics of performance evaluation of semantic algorithms and the implementation of formal models of representation of the domain knowledge in the form of ontological and developed algorithms in software system terminology extraction from text are considered.

Introduction

Operation principles of the existing term extraction algorithms in lexicography and terminology are based on statistical and linguistic methods.

When analysing big documentation corpora, the domain specific nature is essential to be taken into consideration to obtain a list of terms specific for this domain. Semantic algorithms based on the determined semantic metrics are used to solve such tasks.

One of the most universal methods to present expert knowledge in the context of semantic description comprehensiveness is the ontological approach. That is why one of the most important ways to solve the problem of terminology extraction from a big corpus of technical documentation is development and use of semantic metrics based on ontological models.

Main part

The operation algorithm of the term extraction system developed supposes the following execution sequence:

1. Processing a text via the morphological analysis module
2. Calculation of linguistic and statistical characteristics of the text processed and containing morphological tagging via the main system module
3. Calculation of semantic characteristics of words and word combinations of the processed text based on the presented methods using the OWL ontology developed

We used a text of 62,000 words that is a part of an operators manual for a CNC turning milling machine to carry out an experiment. To estimate effectiveness of the indicators calculation, we considered such criteria as *Precision*, *Recall* and F_1 for each indicator in each linguistic usage category. The results of the experiment on term extraction from the operators manual for a CNC turning milling machine using the developed ontology of the relevant domain enable us to conclude that nested bond criterion is high efficient to solve the set task, especially in case of tri- and quadri-words.

Conclusion

Thus, the «term/non-term» semantic metric proposed in this paper based on a domain ontology enables us to extract only the terms and term combinations relevant to the domain from an incoming corpus of single- or multi-words establishing a numerical value of relevance to the domain terms for each of the input ones.