КАЧЕСТВЕННАЯ И КОЛИЧЕСТВЕННАЯ ОЦЕНКА ЗНАНИЙ В АКТИВНЫХ СИСТЕМАХ

Выхованец В.С.
(*Институт проблем управления РАН, Москва*)
valery@vykhovanets.ru
Круппа З.П.
(*МГТУ им. Н.Э. Баумана, Москва*)
krouppa@mail.ru

Предложено решение задачи качественной оценки знаний агентов на основе синтаксического анализа текстов, выявления предикативной структуры предложений и построения семантических сетей. Для подсчета объема знаний в семантической сети определено семантическое расстояние. Показано, что объем знаний, содержащихся в семантической сети, является мерой на множестве семантических сетей, а введенное расстояние превращает это множество в метрическое пространство.

Ключевые слова: синтаксический анализ текстов, предикативная структура предложений, семантические сети, семантическое расстояние, объем знаний.

Введение

В связи с усложнением современных активных систем актуальной стала задача количественной и качественной оценки знаний, выражаемых их активными элементами. Например, при рефлексивном (информационном) управлении в реальных социальных системах требуется получение оценок информированности агентов по тем или иным вопросам [4].

Основной формой представления знаний является текст. Трудности, с которыми столкнулись исследователи при извлечении знаний из текста, состоят в том, что до сих пор не решена задача семантического анализа текста, заключающаяся в получении смысла, содержащегося в тексте, и его преобразование в одну из известных форм представления знаний [1, 5].

Однако для многих прикладных областей достаточным оказывается не полное извлечение знаний из текстов, а их количественная и качественная оценка [3].

1. Синтаксис предложений

Простое предложение русского языка имеет предикативную структуру и может быть представлено грамматическим предикатом, аргументами которого являются грамматический субъект и грамматический объект.

Обычно подлежащее выражает грамматический объект, сказуемое – грамматический предикат, а дополнение – грамматический субъект. В свою очередь сложное предложение состоит из простых и имеет в своем составе две или несколько предикативных единиц, образующих в смысловом, конструктивном и интонационном отношении единое целое [2].

Будем предполагать, что грамматический объект и грамматический субъект предложения описывают понятия предметной области, а грамматический предикат – связь между ними. Тогда с помощью синтаксического анализа простого предложения можно выделить одно суждение, которое сообщает о взаимосвязанных понятиях и о характере их взаимосвязи.

2. Семантическая сеть

Пусть заданы две семантические сети: сеть текста S и сеть простого предложения S'. Сеть текста S зададим в виде упорядоченного множества из трех элементов:

(1) $S = (N, E, P)$,

где $N = \{n_i \mid i = 1, \ldots, Q\}$ – множество узлов сети с числом элементов Q, $E = \{(n_i, n_j, p_k) \mid n_i, n_j \in N; p_k \in P\}$ – множество ее дуг, P – множество двуместных предикатов. Дуги заданы упорядоченными множествами из трех элементов $(n_i, n_j, p_k) \in N \times N \times P$, где $n_i \in N$ – началь-

ный узел, $n_j \in N$ – конечный узел, $p_k \in P$ – имя дуги, а × – знак операции декартового произведения множеств.

В свою очередь сеть предложения S' простая и состоит из двух узлов n_1, n_2 и одной дуги, помеченной именем некоторого предиката p:

$S' = (N', E', P')$, $N' = \{n_1, n_2\}$, $E' = \{(n_1, n_2, p)\}$.

Тогда объединением сетей $S = (N, E, P)$ и $S' = (N', E', P')$ будет сеть $S'' = S \cup S'$ такая, что $S'' = (N'', E'', P'')$ и $N'' = N \cup N'$, $E'' = E \cup E'$, $P'' = P \cup P'$.

Аналогично вводится пересечение и разность сетей.

3. Семантическое расстояние

Пусть задана семантическая сеть S. Зафиксируем два произвольных ее узла n_i и n_j. Найдем $R(n_i, n_j)$ – множество путей без циклов (цепей) длины не более чем M, ведущих от узла n_i к узлу n_j. Тогда семантическое расстояние L между узлами n_i и n_j может быть вычислено по формуле:

$$(2) \quad L(n_i, n_j) = \sum_{r \in R(n_i, n_j)} \frac{\min(w_1^r, w_2^r, \ldots, w_{d(r)}^r)}{d(r)},$$

где $d(r)$ – длина пути r, $d(r) \le M$; M – глубина связи; w_i^r – вес дуги i пути r, $i = \overline{1, d(r)}$, min – функция, возвращающая минимальное значение ее аргументов.

Из формулы (2) следует, что два узла отдалены друг от друга, если между ними имеется много путей (понятия слабо связаны). Отдаленность двух узлов тем больше, чем больше веса соединяющих их дуг (более вариативными являются связи между понятиями). Однако если в пути встречается дуга с небольшим весом, то этот путь вносит меньший вклад в удаленность узлов друг от друга. Но не все пути учитываются при подсчете расстояния между узлами: исключаются те пути, длина которых больше заданной глубины связи (трудно установить связь между понятиями, так как это требует использования большого числа предложений).

4. Измерение знаний

Под объемом знаний, содержащихся в семантической сети $S = (N, E, P)$, будем понимать величину, вычисляемую по следующей формуле:

$$(3) \quad K(S) = \sum_{n_i, n_j \in N} L(n_i, n_j),$$

где $K(S)$ – объем знаний в семантической сети S, а $L(n_i, n_j)$ – семантическое расстояние между узлами n_i и n_j, вычисляемое по формуле (2).

Формула (3) утверждает, что объем знаний в сети S есть сумма семантических расстояний между всеми парами ее узлов. Как и у семантического расстояния, единицей измерения объема знаний является грамматический предикат.

Теорема 1. Объем знаний (3) *является аддитивной мерой на множестве семантических сетей.*

Семантическое расстояние между сетями S_1 и S_2 определим как объем знаний, содержащийся в симметрической разности этих сетей:

$$(4) \quad D(S_1, S_2) = K(S_1 \setminus S_2 \cup S_2 \setminus S_1).$$

Теорема 2. Семантическое расстояние (4) *является метрикой на множестве семантических сетей.*

Таким образом, множество семантических сетей текстов является метрическим пространством, а семантическое расстояние между двумя сетями равно суммарному объему знаний, в них содержащихся.

5. Оценка знаний

Количественная и качественная оценка знаний осуществляется путем построения семантических сетей текстов, написанных разными агентами, и сравнения этих семантических сетей.

Пусть имеются следующие семантические сети:

– S – семантическая сеть центра управления;

– S_i ($i = 1, 2, ..., k$) – семантические сети агентов.

Тогда информированность агентов определяется так:

(5) $Y_i = K(S_i \setminus S)$, $y_i = K(S_i \setminus S) / K(S)$ ($i = 1, 2, ..., k$),

где Y_i – абсолютная информированность агента i, а y_i – его относительная информированность. Для качественной оценки информированности i-го агента может использоваться семантическая сеть $S_i \setminus S$. Аналогично вводятся оценки информированности агентов относительно друг друга.

Заключение

В отличие от других известных методов определения объемов знаний, основанных на использовании онтологий и тезаурусов, разработанный метод не привязан к конкретной предметной области и не требует привлечения экспертов для ее первичного описания.

Литература

1. БЕЛОНОГОВ Г.Г. *Компьютерная лингвистика и перспективные информационные технологии* // Русский мир. – 2004. – 248 с.
2. ВАЛГИНА Н.С., РОЗЕНТАЛЬ Д.Э., ФОМИНА М.И. Современный русский язык: Учебник / Под ред. Н.С. Валгиной. – М.: Логос, 2002. – 528 с.
3. НАУМОВ И.С., ВЫХОВАНЕЦ В.С. *Оценка трудности и сложности учебных задач на основе синтаксического анализа текстов* // Управление большими системами. – 2014. – Вып. 48. – С. 97-131.
4. НОВИКОВ Д.Д. *Теория управления организационными системами*. – М.: МПСИ, 2005. – 584 с.
5. ПОПОВ Э.В. *Общение с ЭВМ на естественном языке*. – М.: Наука, 1982. – 360 с.