К вопросу об универсальном представлении концептуальных структур в системах индексирования и автоматической рубрикации текстов

В последние годы растет интерес исследователей к задаче так называемого семантического (интеллектуального) поиска – информационного поиска по содержанию поискового запроса, а не по ключевым словам, из которых он состоит. На сегодняшний день известны некоторые частные решения этой задачи: использование лингвистических тезаурусов и онтологий позволяет осуществлять поиск с учетом отношений семантической эквивалентности (синонимии) и родо-видовых отношений (гиперонимии, отношений наследования). При этом результаты поиска могут оцениваться по степени семантического соответствия поисковому запросу, точнее – образа документа образу запроса. Образ документа – это «Запись основного содержания документа на ИПЯ[1]» [Панков, Захаров 1996: 337]; образ запроса – «... представление запроса на ИПЯ» [Там же].

Вместе с тем, все большую актуальность приобретает задача эффективной автоматической рубрикации. Эта задача состоит в том, чтобы установить степень соответствия текста той или иной рубрике, точнее, опять же, образа документа образу рубрики. Образ рубрики – это логическая формула, в действительности по структуре сходная с формулой сложного поискового запроса (см. [Добров 2011]).

При индексировании и при рубрикации в тексте выявляются одни и те же единицы, над которыми производятся весьма сходные операции, что позволяет предположить возможность применения универсального способа моделирования этих единиц, оптимального для обеих этих задач.

В области информационного поиска была предложена идея дескрипторов – «слов, являющихся именами простых понятий», которые «...

1 ИПЯ – информационно-поисковый язык; его основные характеристики существенным образом зависят от функционального назначения информационно-поисковой системы

выступают в качестве как бы координат документов в некоторм умозрительном n-мерном предметно-тематическом пространстве» [Панков, Захаров 1996: 343]. В случае семантического поиска или автоматической рубрикации с учетом семантических отношений представляется не вполне очевидным, чему конкретно должны соответствовать дескрипторы: лексическим единицам (ЛЕ), их значениям, или более крупным единицам.

Если дескрипторы соответствуют только ЛЕ, то возникает необходимость учета морфологической неоднозначности (так, словоформе «*правила*» соответствуют две ЛЕ – «*править*» и «*правило*», а, значит, и два дескриптора), а лексическая неоднозначность не учитывается вообще (так, словоформа «*орган*» может соответствовать двум омонимичным ЛЕ – «*о́рган*» и «*орга́н*», причем первая единица может обозначать, по крайней мере, 'часть организма' или 'элемент управленченской структуры', а вторая – 'музыкальный инструмент'); в результате не может работать поиск или рубрикация с учетом семантических отношений, устанавливающихся не между ЛЕ, а между их значениями.

Если дескрипторы соответствуют значениям ЛЕ (каждому значению соответствует отдельный дескриптор), то, помимо необходимости учета морфологической неоднозначности возникает необходимость учета лексической неоднозначности. Кроме того, в обоих случаях возникает проблема так называемой ложной корреляции (см. [Агеев, Добров, Лукашевич 2008: 25]): при поиске по запросу «*экономическая реформа*» или при рубрикации по аналогичной такому запросу рубрике выделяются тексты, в которых содержатся словоформы имени прилагательного «*экономический*» и имени существительного «*реформа*», но в действительности речь идет, например, об экономических последствиях реформы в системе образования.

Таким образом, для эффективного семантического поиска и автоматической рубрикации документов с учетом семантических отношений каждый дескриптор должен соответствовать значению любого целостного

словосочетания или предложения, которое предполагается возможным найти по запросу или по элементу образа рубрики.

В отличие от множества ЛЕ и их значений, множество значений всех возможных синтаксических единиц принципиально бесконечно и не может быть сопоставлено дескрипторам заранее. Поэтому база данных, в которой устанавливаются соответствия между значениями синтаксических единиц и дескрипторами, должна быть динамической и непрерывно пополняться при индексировании и автоматической рубрикации.

Для построения такой базы данных необходимо реализовать универсальное представление значений синтаксических единиц, способное, по крайней мере, работать в функции поискового ключа в базе данных. Эта задача не имеет однозначного решения, поскольку на сегодняшний день в лингвистике нет общепринятого ответа на вопрос, что именно представляет собой значение словосочетания или предложения и какая именно математическая модель может быть использована для его представления. Некоторые исследователи ([Адамец 1978], [Арутюнова 1976], [Лакофф 1981], [Макколи 1981], [Падучева 1974] и др.) используют полипропозициональные структуры (логические формулы), однако эти структуры не могут быть единообразными с точки зрения сопоставления тем или иным аргументам пропозиций их ролей в этих пропозициях. Указанная проблема решается путем осложнения полипропозициональных структур указанием глубинных ролей ([Апресян 1995: 126, 127], [Богданов 1996], [Филлмор 1981а], [Филлмор 1981б], [Чейф 1975], [Шенк 1980: 42] и др.), однако при этом возникает проблема отсутствия общепринятого инвентаря этих ролей и правил их использования. Многие исследователи вообще не используют древовидные пропозициональные структуры, пользуясь вместо них сетевыми моделями (семантическое представление в модели «Смысл ⇔ Текст» [Мельчук 1974], концептуальные графы J. Sowa [Sowa 1976]).

Использование графов для представления значений синтаксических

единиц оправдано тем, что значение любой неидиоматической синтаксической единицы можно представить в виде некоторого отношения между значениями ее непосредственных составляющих (НС). Данное утверждение в некоторой степени следует из принципа композициональности (см. [Werning, Machery, Schurz 2004]). При этом направление отношения в каждом случае определяется глубинными ролями, в которых участвуют значения НС, зависящими от направления синтаксической связи и семантических валентностей значений этих НС.

Тем не менее, дуги (ребра) графов могут соединять узлы (вершины), но не могут соединять другие дуги (ребра), однако некоторые ЛЕ сами по себе обозначают отношения между объектами, которые, в свою очередь, могут вступать в отношения друг с другом и с другими объектами. Так, значения глаголов могут рассматриваться, например, как отношения между субъектами и объектами; значения предлогов могут рассматриваться как отношения между значениями глаголов и значениями именных групп, значения союзов могут представлять собой отношения между любыми единицами – как объектами, так и отношениями.

Таким образом, оптимальной структурой для моделирования семантики синтаксических единиц представляется сетевая модель фрактального типа, допускающая возможность использования ребер (дуг) в качестве вершин других ребер (дуг). Эта модель может быть определена следующим образом:

NF (фрактальная сеть) = <O,R,E>, где

O – непустое множество объектов ($|O| > 0$);

R – множество отношений, подмножество O, ($|R| > 0$, $R \subseteq O$);

E – множество дуг (ребер).

При этом каждая дуга – это тройка «объект, отношение, объект» ($E = \{<s,r,o>: s \in O, r \in R, o \in O\}$); все объекты, не являющиеся отношениями,

являются началами или концами хотя бы одной дуги ($\forall x, x \in O \setminus R: \exists \langle s,r,o \rangle \in E, (x = s) \vee (x = o)$); все отношения представлены хотя бы одной дугой ($\forall x, x \in R: \exists \langle s,r,o \rangle \in E, x=r$).

Вышеуказанная структура предоставляет возможность моделировать одни и те же значения синтаксических единиц множеством различных способов. Например, связь между субъектом и предикатом может рассматриваться как элемент множества R (отношение 'участие в роли субъекта действия'), в то же время и сам предикат может рассматриваться как элемент этого множества, поскольку связывает субъект и объект; наконец, между субъектом и его связью с предикатом можно также выделить некоторую связь ('участие в роли субъекта отношения «участие в роли субъекта действия»') и т.д. (см. рис. 1).

В связи с данным обстоятельстом возникает вопрос о введении некоторых дополнительных ограничений, которые могли бы позволить технически организовать представление этой структуры таким образом, чтобы ее можно было использовать в виде уникального ключа.

Представляется очевидным, что указание отношения между объектом и уже имеющимся отношением, в котором участвует этот объект, избыточно по своей сути. Такое ограничение можно сформулировать следующим образом:

$\forall \langle s,r,o \rangle \in E: \forall r' \in R:$

$\langle s,r',r \rangle \notin E \wedge \langle r,r',s \rangle \notin E \wedge \langle o,r',r \rangle \notin E \wedge \langle r,r',o \rangle \notin E.$

Рис. 1: Возможные отношения между субъектом, предикатом и объектом

Указанное правило позволяет нормировать различные представления одних и тех же семантических структур, сводя их к минимальным по структурной сложности.

Представляется также очевидным, что изображение в виде узлов тех элементов множества O, которые входят во множество R, не менее избыточно. Такое изображение может потребоваться при представлении семантической структуры в виде графа, не допускающего возможности участия ребер (дуг) в роли вершин, но не в случае фрактальной сети. Таким образом, например, семантическая структура предложения «*Мама мыла раму*», даже с учетом семантики прошедшего времени, приобретает весьма компактный вид (см. рис. 2).

Рис. 2: Семантическая структура предложения "Мама мыла раму"

Данная математическая моель может быть представлена в машинной памяти в виде массива структур вида <s,r,o,e>, где s – номер первого элемента

массива, описывающего объект, являющийся началом ребра, r – номер первого элемента массива, описывающего отношение, соответствующее этому ребру, o – номер первого элемента массива, описывающего объект, являющийся концом ребра, или идентификатор концепта онтологии или элемента тезауруса, соответствующего концу ребра; e – булево поле, принимающее значение «истина», если o указывает на элемент массива, или «ложь», если o указывает на внешнюю по отношению к этому массиву онтологию или тезаурус.

В таблице 1 представлено содержимое массива, соответствующего значению предложения «*Мама мыла раму*», в таблице 2 – соответствующий этому массиву фрагмент тезауруса.

номер в массиве	**s**	**r**	**o**	**e**
0	0	2	6546468	ЛОЖЬ
1	0	3	5	ИСТИНА
2	2	2	1151563	ЛОЖЬ
3	3	2	6523643	ЛОЖЬ
4	3	6	1256476	ЛОЖЬ
5	5	2	7687657	ЛОЖЬ
6	6	2	2765786	ЛОЖЬ

Таблица 1: Массив структур, соответствующий значению предложения "Мама мыла раму"

Идентификатор в тезаурусе/онтологии	**Концепт онтологии / элемент тезауруса**
1151563	Наследование (отношение между видом и родом)
1256476	Момент речи (момент времени, в который было произведено высказывание)
2765786	До (предшествование во времени)
6523643	Мыть (осуществлять чистку при помощи жидкости)

6546468	Мама (родитель женского пола)
7687657	Рама (конструкция, содержащая жесткие связи между элементами)

Таблица 2: Элементы тезауруса/онтологии, на которые ссылаются элементы таблицы 1

Представление семантики синтаксических единиц в виде массивов, т.е. непрерывных областей машинной памяти, может быть использовано в качестве ключа в базе данных. Указанный способ представления реализован автором настоящей статьи в системе автоматической обработки текстов AIRE[2], используемой в действующей системе семантического поиска и автоматической рубрикации текстов MinIRE[3]. Обе системы являются свободным программным обеспечением, опубликованным и распространяемым под лицензией GNU GPL.

Литература

1. Агеев М.С., Добров Б.В., Лукашевич Н.В. Автоматическая рубрикация текстов: методы и проблемы // Учебные записки Казанского Государственного Университета. Т. 150, кн. 4 — 2008
2. Адамец П. Образование предложений из пропозиций. Прага, 1978
3. Апресян Ю.Д. Избранные труды, том I. Лексическая семантика. Синонимические средства языка. – М.: Школа "Языки русской культуры", Издательская фирма "Восточная литература" РАН, 1995
4. Арутюнова Н.Д. Понятие пропозиции в логике и в лингвистике // Изв. АН СССР. Сер. лит. и яз. № 1, Т. 35. — 1976
5. Богданов В.В. Моделирование семантики предложения // Прикладное языкознание: Учебник. — СПб, 1996.
6. Добров А.В. Комплексный лингвистический подход к автоматической рубрикации новостных сообщений // Политическая лингвистика. Вып. 3(37) – Екатеринбург, изд-во УРГПУ: 2011 – сс. 202-209

2 Artificial Intelligence Information Retrieval Engine

3 Minimalist Intellectual Information Retrieval Engine

7. Лакофф Дж. О порождающей семантике // Новое в зарубежной лингвистике. Вып. X: Лингвистическая семантика. — М., 1981

8. Макколи Дж. О месте семантики в грамматике языка // Новое в зарубежной лингвистике. Вып. X: Лингвистическая семантика. — М., 1981

9. Мельчук И.А. Опыт теории лингвистических моделей «Смысл ↔ Текст». М., Наука, 1974

10. Падучева Е.В. О семантике синтаксиса. М, 1974

11. Панков И.П., Захаров В.П. Информационно-поисковые системы // Прикладное языкознание: Учебник. — СПб, 1996.

12. Филлмор Ч. Дело о падеже // Новое в зарубежной лингвистике. Вып. X: Лингвистическая семантика. — М., 1981а

13. Филлмор Ч. Дело о падеже открывается вновь // Новое в зарубежной лингвистике. Вып. X: Лингвистическая семантика. — М., 1981б

14. Чейф У. Значение и структура языка. М.: Прогресс, 1975

15. Шенк Р. Обработка концептуальной информации. Пер. с англ. – М.: «Энергия», 1980

16. Sowa J.F. Conceptual Graphs for a Data Base Interface // IBM Journal of Research and Development 20 (4). – 1976 – pp. 336–357.

17. Werning M., Machery E., Schurz G. The Compositionality of Meaning and Content. Vol. 1, 2 – Piscataway, NJ: "Transaction", White Cross Mills, Lancaser: "Gazelle" – 2004