УДК 004.421.5

Самковский С.К.,
Стиренко С.Г., к.т.н.

ОПТИМИЗАЦИЯ ПРЕДСТАВЛЕНИЯ ДАННЫХ В СЕМАНТИЧЕСКОЙ СЕТИ СО СТРОГОЙ ТИПИЗАЦИЕЙ

**Национальный технический университет Украины
«Киевский политехнический институт»**

Предложена элементная база и правила формирования неоднородной динамической бинарной семантической сети со строгой типизацией, создаваемой для решения задачи семантического анализа предложений на естественном языке.

Введение

Одним из средств представления знаний в предметной области, в частности, семантики текстовой информации, представленной на естественном языке, является семантическая сеть. Однако базовый инструментарий семантической сети, доступный при построении информационной модели, трудно формализуем, и неприменим при ее программной реализации. В статье предлагается описание средств оптимизации элементной базы и принципов построения семантической сети с тем, чтобы сделать ее доступной для программной интерпретации.

Целью работы является разработка элементной базы и набора правил, по которым текст, представленный на естественном языке, можно будет транслировать в семантическую модель без потери смысла.

Подходы к решению задачи

Задача состоит в формировании элементной базы и набора правил, достаточных для программного хранения и обработки семантики предложений на естественном языке. Таким образом, входными данными алгоритма будет набор текстовой информации произвольного содержания, а на выходе должна быть сформирована логическая модель. Затем должен быть реализован алгоритм извлечения предварительно сохраненных знаний, такой, что если на вход подается некоторое слово, упоминаемое ранее и представленное в базе, модель должна содержать в себе механизмы для для извлечения связанной с этим словом информации. Так как семантическая сеть является моделированием ассоциативной природы человеческой памяти и мышления, идеальным решением задачи было бы извлечение информации не по отдельному слову, а по конкретному вопросу, причем в связной форме. Однако такой подход выходит за рамки этой работы.

Построение информационной модели используется при выполнении машинного перевода текста. Большинство существующих алгоритмов перевода в основе имеют идею *UNL* (универсального сетевого языка). Текст на исходном языке транслируется в *UNL*, а затем на целевой язык. Такой двухэтапный перевод с одной стороны увеличивает время выполнения алгоритма, но с другой, при последовательном переводе одного и того же текста на несколько разных языков, ошибка перевода не накапливается, а остается фиксированной. Также в памяти необходимо хранить не $n(n-1)$ словарей, а только $2n$. Теоретически текст, последовательно переведенный на несколько других языков, а затем на исходный, останется полностью неизменным. Но на практике достичь этого не удается, поскольку нельзя каждому из слов в предложении, например, на английском языке, сопоставить некоторое слово *UNL*, а затем в том же порядке воспроизвести все слова из словаря другого языка. Каждый язык имеет свои уникальные правила синтаксического строя предложений, поэтому, чем сложнее предложение, тем более несогласованным получается перевод [1]. Для решения этой проблемы, подобные алгоритмы содержат ряд надстроек, которые формируют некоторые логические схемы из слов в памяти, используя словари синтаксических и орфографических правил для каждого из языков. Идеальной интерпретацией такого алгоритма было бы построение самодостаточной семантической сети в памяти компьютера по исходному предложению, а затем, используя словарь другого языка, интерпретировать эту семантическую сеть. Авторами предлагается усовершенствованная модель логической ин-

терпретации содержания предложений, используемых при машинном переводе текста.

Проблема автоматизации текстового анализа

Рассматриваемая модель оперирует выражениями на естественном языке, при этом смысловой единицей выступает слово. Это означает, что задача носит не только информационно-вычислительный, но и лингвистический характер. Для построения адекватной модели информации необходимы данные о том, какой тип имеет каждое слово, используемое в предложении, и в каких взаимоотношениях находятся эти слова. При этом предполагается разбор грамматической структуры предложения, совместный с орфографическим анализом каждого из слов для выявления их согласованности. Задача разбора грамматической структуры предложений является очень сложной и трудоемкой ввиду слабой формализуемости естественного языка. На данный момент не существует алгоритмов, полностью решающих эти проблемы, однако существует ряд программных решений, возможности которых достаточны для использования их с целью машинного разбора грамматической структуры предложений, необходимого в рамках поставленной задачи [2]. Среди них *Stanford Parser* [3], *Link Grammar* [4], *VISL*, *ASSP* и др. Все они предназначены для выявления структур в фразе и построения диаграммы предложения. Такие диаграммы формируются с использованием формализованных словарей терминов и представляют собой дерево сущностей (слов). Каждая из сущностей в таком дереве имеет определенный тип, набор которых фиксирован. Тип сущности имеет лингвистический смысл, так предложение как правило имеет две части: относящиеся к предмету и действию, описываемых в нем. Часть предложения, относящаяся к предмету имеет тип *NP*, другая, относящаяся к действию – *VP*. В данной работе используется *Stanford Parser*, как один из наиболее мощных инструментов.

Алгоритм построения семантической сети из отрывка текста должен опираться на результаты такого машинного разбора грамматической структуры, то есть оперировать порождаемыми им сущностями. Такой подход к анализу текста считается высокоуровневым, поскольку оперирует такими абстрактными сущностями как отдельные слова. Это существенно упрощает алгоритм, однако делает его зависимым от ошибок, производимых сторонними программами. В перспективе должен быть разработан алгоритм, который, исходя из анализа разнообразных предложений, самостоятельно обобщит правила, реализуемые парсером, что даст возможность не только учесть ошибки имеющихся программ, но и сделать алгоритм разбора предложений на естественном языке обучаемым.

Обоснование выбора типа семантической сети для решения поставленной задачи

Основной целью использования семантической сети является построение информационной модели предметной области с целью дальнейшего извлечения представленной информации и работы с ней. Одним из достоинств семантических сетей является их графическая нотация: они наглядны и удобны для отображения памяти и предикатов описывающих сущности. Задачей является моделирование процесса мышления в виде ориентированного графа, который дает возможность отобразить ассоциативную природу связей между сущностями.

Существует несколько разновидностей семантических сетей, которые отличаются по элементной базе, используемой в представлениях, и законах, по которых сеть формируется. В работе рассматривается динамическая сеть, способная к расширению во время выполнения программы, что накладывает определенные требования на элементную базу: она должна быть фиксированной и не должна изменяться. Семантическая сеть, соответствующая таким условиям будем называть сетью со строгой типизацией. Выбранный набор элементов и отношений должен быть достаточным для того, чтобы отображать смысловые отношения, представленные в тексте. Для упрощения, была выбрана модель с бинарными отношениями, то есть такая, где одна связь (*link*) связывает не более двух элементов сети. Поскольку связи между элементами семантической сети являются прототипами нейронных синоптических связей в мозгу человека, такой постулат далек от истины, однако в рамках поставленной задачи использование n-арных связей неоправданно усложнило бы задачу. Их применение показывает положительные результаты в сетях без строгой типизации, поскольку задачи, которые ставятся перед ними, отличаются от тех, которые ставят-

ся перед рассматриваемыми сетями со строгой типизацией. Семантические сети без строгой типизации концентрируются на последовательностях подаваемых на их вход сигналов, выделяя паттерны в них и формируя сущности паттернов. Сети со строгой типизацией – более абстрактные, поскольку работают уже со сформированными грамматически связанными понятиями.

Построение подобной модели требует не только определения элементной базы, но и разработки четких алгоритмов развития семантической сети. Ее развитие состоит в росте и происходит за счет поступления новой информации. Такой подход требует наличия четких алгоритмов по расширению существующей сети за счет информации, представленной в текстовом виде на естественном языке, а также, алгоритмы считывания (добычи) информации. Таким образом, для решения задачи будет использоваться неоднородная динамическая бинарная семантическая сеть со строгой типизацией [5].

Выбор элементной базы семантической сети

Поскольку задача разбора грамматической структуры предложений была реализована в сторонней библиотеке Stanford Parser, элементная база семантической сети со строгой типизацией должна коррелироваться с элементной базой, используемой при разборе предложений. Поскольку семантическая сеть динамическая, набор элементов должен быть строго фиксирован и неизменен. Расширение семантической сети должно производиться за счет расширения количества запоминаемых элементов, а не изменения их свойств.

Двумя базовыми элементами семантической сети со строгой типизацией выбрано три: сущность (*Entity*), действие (*Action*) и качество (*Quality*), которые соответствуют существительному, глаголу и прилагательному при синтаксическом разборе. Все остальные части речи представляются производными от этих трех базовых. На морфологическом уровне мозг человека оперирует такими же сущностями – *Entity*, *Action* & *Quality* [6].

Предполагается иерархическая и в то же время ассоциативная структура формируемой семантической сети: каждая сущность может иметь произвольное количество разрешенного типа связей с другими сущностями, при этом семантика связей кодируется через типы используемых элементов и вид их соединений.

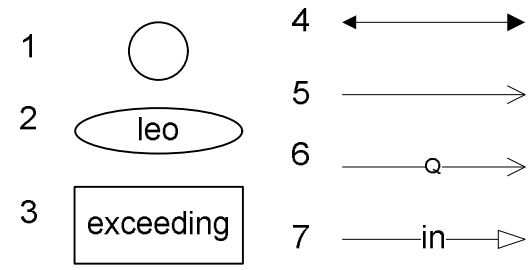

Рис.1. Элементная база семантической сети со строгой типизацией

Всего при построении информационной модели будет использоваться 3 типа сущностей и 4 типа связей между ними (рис.1). Обозначение **2** представляет *Entity* и *Quality*. Объединение обозначений вызвано тем, что отличия между ними определяются их использованием, а не их принадлежностью к определенному классу частей речи. Одна и та же сущность может выступать как *Entity* в одном предложении и как *Quality* в другом. Поэтому это отличие было представлено с помощью специального типа связи, а не элемента. Свойством фигуры 2 является имя, которым является слово, представляющее эту сущность. Обозначение **1** – это подсущность. Она используется для построения древовидных структур и синтезируется из нескольких фигур. Подсущность также как и сущность может быть описана именем, но для его получения необходим специальный алгоритм, который соберет это имя из имен других сущностей, соотносящихся с этой подсущностью. По существу, это ключевой элемент модели, поскольку с его помощью возможна реализация алгоритмов абстракции и конкретизации. Фигура **3** используется для обозначения действия (*Action*). Точно так же как и сущность, может обладать именем (соответствующим глаголом), или, как подсущность, являться синтезом нескольких действий (*Action*).

Используемые в семантической сети связями являются обозначения 4-7. Связь **4** представляет собой связь эквивалентности, **5** – принадлежности. Связь **6** используется для указания того, что одна сущность характеризуется другой (*Quality*). Связь 7 – это обозначение предлога. Она может характеризоваться именем (*in*, *at*, *with*, *on*, *etc*.), которое будет указывать, с каким условием одна сущность соотносится с другой. Количество этих имен лимитировано и фиксировано. Должен быть разработан специальный алгоритм выявления

такого рода слов для расширения этого списка, однако на данном этапе предлагается использование словаря предлогов. В целом, базовая редакция модели семантической сети со строгой типизацией предполагает правило, согласно которому свойства связей в модели носят статический и унифицированный характер. То есть использование варьирующих имен на связях исключается. Действительно, как будет показано позже, тех же результатов кодирования информации можно достичь и обходясь без такого типа связи как 7, однако это неоправданно усложнит как саму графическую модель, так и ее восприятие, поэтому в данной работе рассматриваются два варианта элементных баз семантической сети со строгой типизацией.

Определение поведения элементов в семантической сети со строгой типизацией.

Модель семантической сети со строгой типизацией предполагает не только четкое определение элементов, используемых в сети, но и правил их взаимного размещения и комбинации. Предполагается, что набор правил, как и элементная база носят статический характер и не расширяется в процессе работы модели.

На рис.2 показаны функциональные свойства элемента *Entity*. Они одинаковы для сущности и подсущности. *Entity* может быть подэлементом другой *Entity* или включать в себя некоторый произвольный набор *Entity*. Она может быть эквивалентной другой *Entity*, а также, характеризоваться некоторым набором *Quality*. Также в структурных отношениях с другими *Entity* могут присутствовать условия выраженные с помощью предлогов (*preposition*).

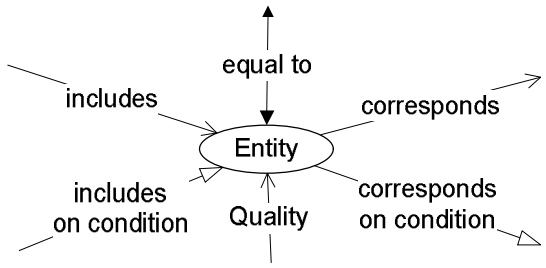

Рис.2. Функциональные свойства элемента Entity в модели семантической сети со строгой типизацией

Рис.3 показывает поведение элемента *Action*. Основными функциональными особенностями действия является то, что всегда у него есть Entity, выполняющее его, и ряд *Entity*, которые на некоторых условиях задействованы в нем. Например, в выражении «Пес бежал по улице», у действия «бежал» актором (который *does* на схеме) будет пес, а функтором (который *uses*) будет улица. Условие – «по», которое будет отображено с помощью связи типа предлог.

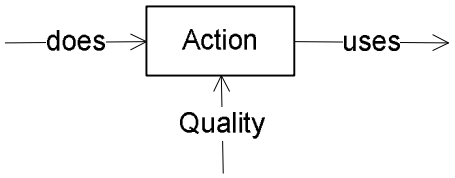

Рис.3. Функциональные свойства элемента Action в модели семантической сети со строгой типизацией

Самым простым поведением в модели обладает свойство, *Quality*, изображенное на рис.4. *Quality* может характеризовать *Entity* или *Action*, а также, само может характеризоваться другим свойством. Например, в выражении «невероятно красивая картина», слово «красивая» является свойством слова «картина», а «невероятно» характеризует слово «красивая», что может быть представлено в модели семантической сети со строгой типизацией с помощью иерархической зависимости *Quality* элементов.

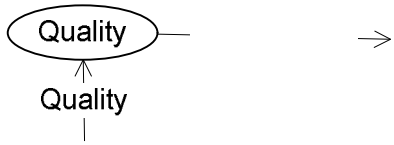

Рис.4. Функциональные свойства элемента Quality в модели семантической сети со строгой типизацией

Пример использования семантической сети со строгой типизацией

Для тестирования модели было выбрано два сложноподчиненных предложения из области естествознания. Поскольку *Stanford Parser* работает только с английским языком, предложения были взяты на нем.

The lion (Panthera leo) is one of the four big cats in the genus Panthera, and a member of the family Felidae. With some males exceeding 250 kg in weight, it is the second-largest living cat after the tiger.

На рис. 5 и 6 представлены результаты работы *Stanford Parser*. Построенная семантическая сеть со строгой типизацией, согласно описанным выше правилам, должна соблю-

дать иерархию синтаксических сущностей результата разбора, представленного *Stanford Parser*.

На рис. 7 представлено результат построения такой модели с использованием связей типа предлог (preposition, *PP* на схеме разбора предложений с помощью *Stanford Parser*). Такая интерпретация знаний представляет собой удобную графическую форму, которую легко можно считать.

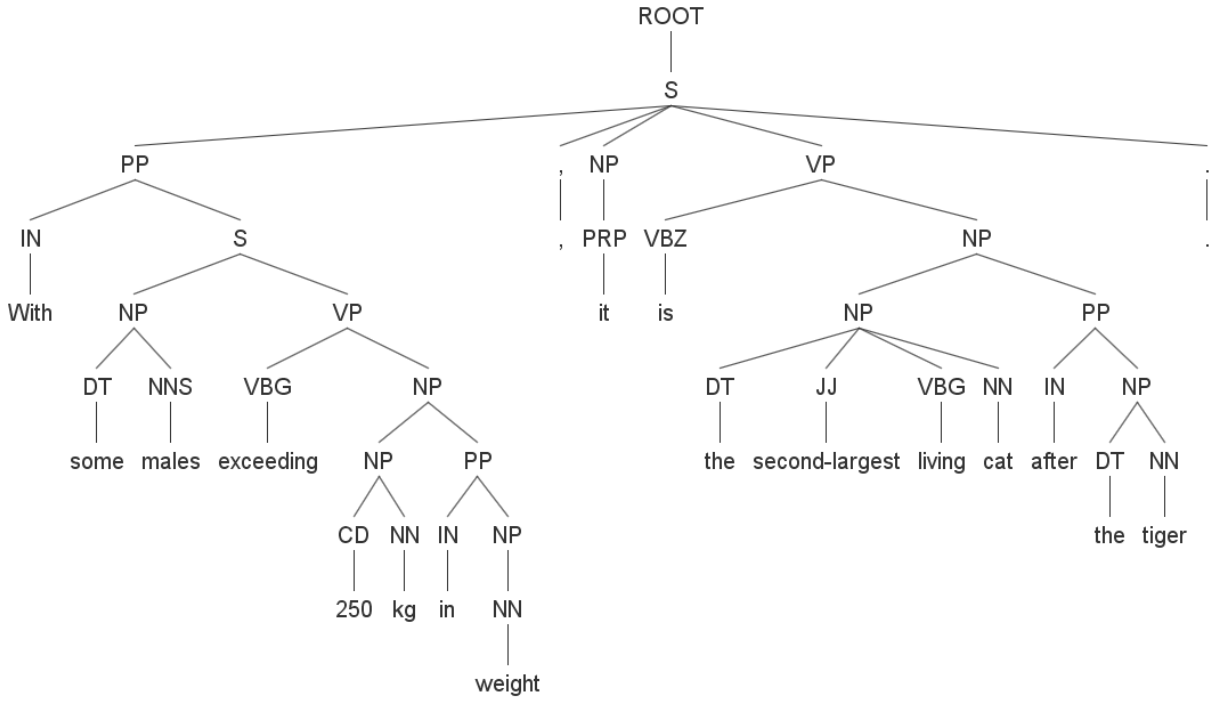

Рис.5. Разбор грамматической структуры предложения с помощью *Stanford Parser*

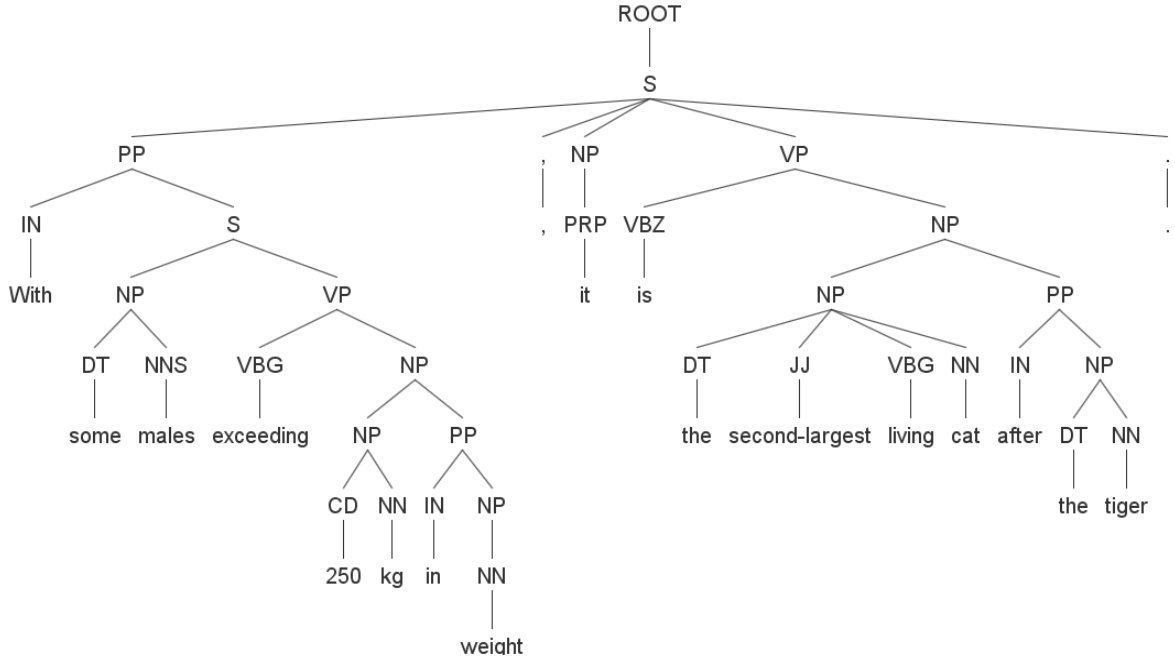

Рис.6. Разбор грамматической структуры предложения с помощью *Stanford Parser*

В одной модели скомпонованы оба предложения. Следует отметить, что для реализации такой компоновки необходим дополнительный алгоритм связи между предложениями. Так, «*it*» во втором предложении указывает на «*lion*» в первом. Однако это неочевидно и требует дополнительного анализа.

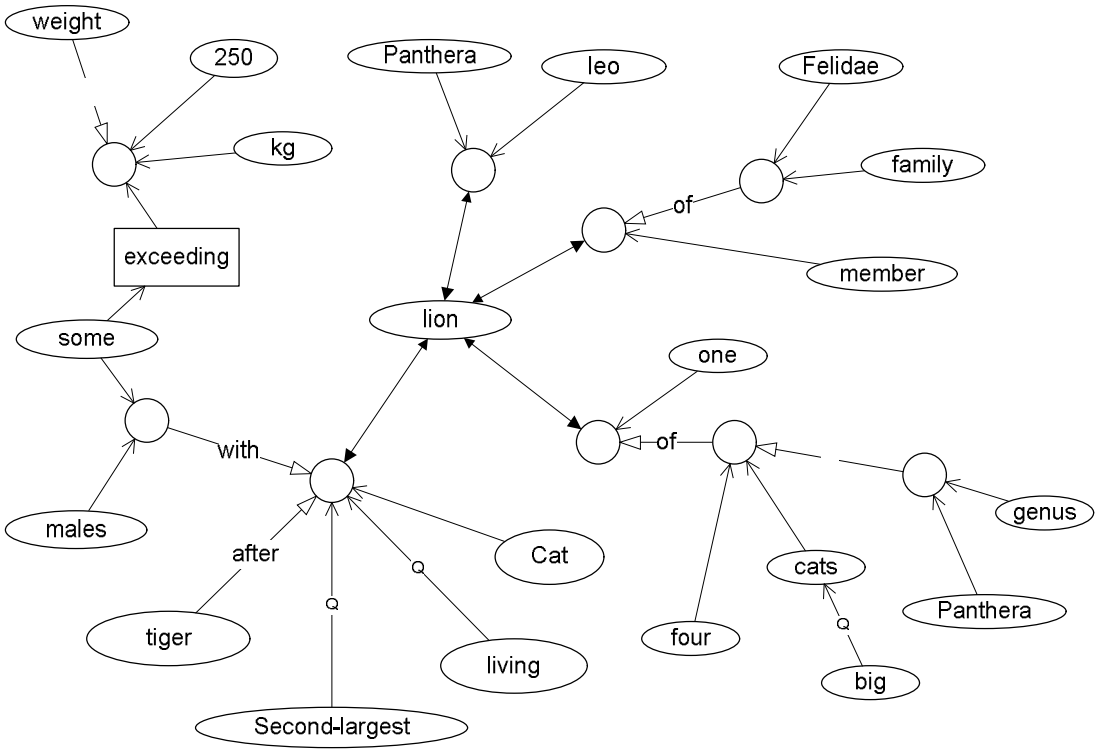

Рис.7. Семантическая сеть со строгой типизацией, описывающая два предложения

Анализ полученных результатов

Основная идея построения такого рода моделей состоит в том, чтобы они давали возможность извлекать фрагменты информации позже. Ведь для 100% воспроизведения информации достаточно просто запомнить исходные предложения. Однако обычно задача состоит в том, чтобы даже с допустимой потерей информации, установить связь между разрозненными кусками информации и синтезировать некоторый обобщенный набор сведений, представленный в удобной для чтения форме.

Так например, при необходимости узнать информацию, релевантную со словом «*Felidae*», применительно к этой модели семантической сети со строгой типизацией, выполнится активация вершины «*Felidae*» и анализ ее связей, что приведет к тому, что существует такая сущность как «*Felidae family*», и «*lion*» – «*member of Felidae family*», а затем, ряд другой релевантной с сущностью «*lion*» информации.

Как описывалось ранее, подобная модель семантической сети со строгой типизацией может обойтись без связей с динамически изменяющимися свойствами (как связь типа предлог). Такая связь может быть заменена абстракцией с помощью дополнительного элемента типа подсущность, которая будет отображать собой условие, на котором определенные Entity связываются с другими элементами семантической сети. Пример такой реализации представлен на рис.8.

Выводы

На примере двух сложноподчиненных предложений было продемонстрированы результаты построения модели неоднородной динамической бинарной семантической сети со строгой типизацией с использованием разбора грамматической структуры предложений, *Stanford Parser*. Были определены элементная база используемая для построения таких моделей и правила, по которым она строится. Основным преимуществом такой модели является ее структурная корреляция с существующими моделями структуры предложений (*phrase structure rules*), что делает удобной ее программную имплементацию.

Семантическая сеть со строгой типизацией является некоторым аналогом нейронной сети, формируемой в кортексе коры головного мозга с рядом очень грубых допущений, призванных сделать ее удобной для решения задачи семантического анализа предложений на естественном языке. Далее планируется разработка алгоритма семантической сети без строгой типизации, который бы создавал классы сущностей самостоятельно, как результат обучения. Также в перспективе необ-

ходимо классифицировать алгоритмы извлечения связной информации из семантической сети и постпроцессинга представленной в ней информации.

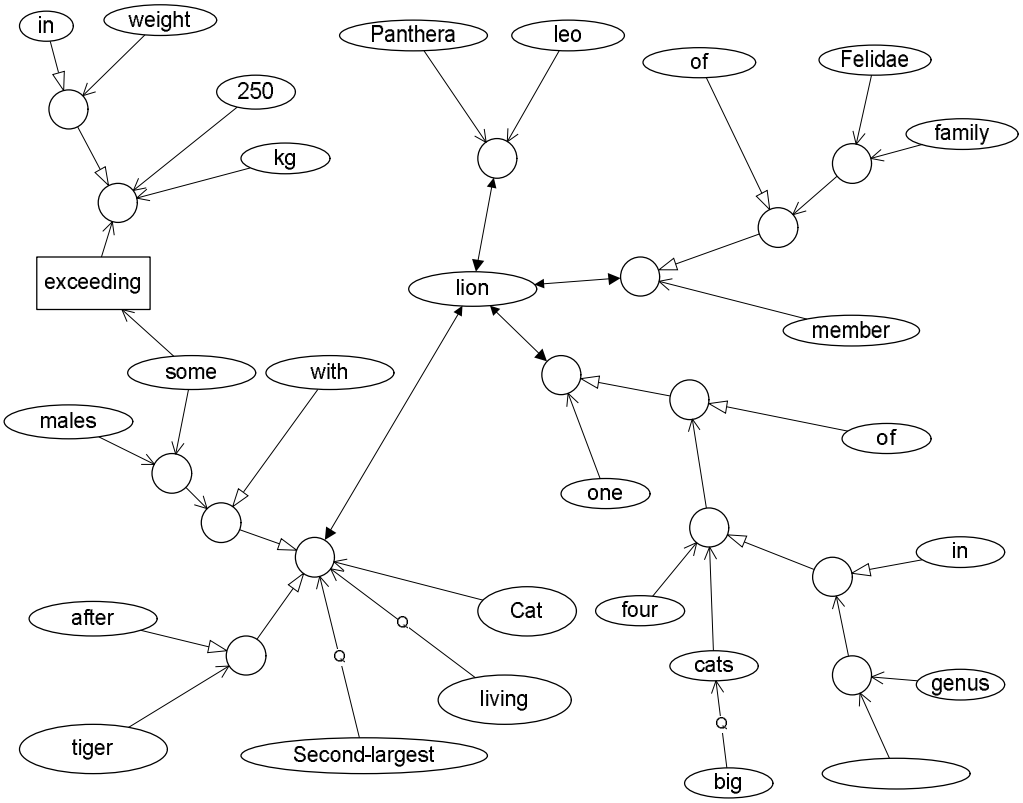

Рис.8. Семантическая сеть со строгой типизацией, описывающая два предложения без использования связей типа предлог

Список литературы

1. Gelbukh A. Universal Networking Language: Advances in Theory and Applications. – Mexico.: Instituto Politécnico Nacional, 2005. – P. 77–101

2. Mulroy D. The War Against Grammar. – Portsmouth.: Heinemann, 2003. – P. 120–124

3. http://nlp.stanford.edu/software/lex-parser.shtml

4. http://www.link.cs.cmu.edu/link/dict/index.html

5. Jurafsky D., Martin J. Speech and Language Processing. – New Jersey.: Pearson, 2009. – P. 427–459.

6. Хьюбел Д. Глаз, мозг, зрение. – М.: Мир, 1990. – С. 47–49.