Д.Р. Рахимова
*Казахский национальный университет им. аль-Фараби, Алматы, Казахстан;*
*e-mail: Diana.Rakhimova@kaznu.kz*

**Построение семантических отношении в машинном переводе**

В данной статье рассматривается классификация концептуальных объектов текста и семантические отношения. Предложены виды и метод описания семантических отношений естественных языков для прикладных лингвистических задач, как автоматическая обработка текста и машинный перевод. Разработан алгоритм построения семантических отношений. Для проверки эффективности описываемый метод применен в реализации машинного перевода различных языковых групп, как русский и казахский язык. Приведены практические результаты.
Ключевые слова: семантические отношения, машинный перевод, алгоритм, русский язык, казахский язык.

**Типы и способы представления семантических отношений.**

Наиболее распространенным способом графического представления семантического отношения (СО) между значениями слов является представление его в виде направленной дуги или стрелки, связывающей между собой точки, или узлы, соответствующие значениям слов. наглядным примером может быть пример семантической сети проиллюстрирована на рисунке 1.


Рис.1. Пример семантической сети

Каждое слово в языке характеризуется определенным набором семантических отношений, в которые оно может вступать с другими словами в тексте.

Количество типов отношений в **семантической сети** определяется её создателем, исходя из конкретных целей. В реальном мире их число стремится к бесконечности. Каждое отношение является, по сути, предикатом, простым или составным. Скорость работы с базой знаний зависит от того, насколько эффективно реализованы программы обработки нужных отношений. [3]

В разных системах формально-семантического описания выделяются разные наборы СО. Так, в **падежной грамматике Ч. Филлмора** выделяется 6 СО, называемых «глубинными падежами»:

1. Агентив (А) — падеж одушевленного инициатора действия.
2. Инструменталис (I) — падеж неодушевленной силы или предмета, который включен в действие или состояние, называемое глаголом, в качестве его причины.

3. Датив (D) — падеж одушевленного существа, которое затрагивается состоянием или действием, называемым глаголом.

4. Фактитив (F) — падеж предмета / существа, который возникает в результате действия или состояния, называемого глаголом.

5. Локатив (L) — местоположение или пространственная ориентация действия или состояния, называемого глаголом.

6. Объектив (O) — семантически наиболее нейтральный падеж: что-либо, что может быть обозначено существительным, роль которого в действии или состоянии, которое называет глагол, определяется интерпретацией самого глагола.

В языке для описания значений слов, предложенном **Ю. Д. Апресяном** между семантическими единицами (семами) устанавливается всего восемь элементарных СО (ЭСО): субъекта, объекта, контрагента, содержания, места, времени, количества и определительное.

**А .Сокиркой** были предложены 25 видов семантических отношении, используемые в модуле поверхностно семантического анализа в системе «Диалинг» для русского языка. [4]

Кроме перечисленных отношений, в программе используются еще некоторых «технические» связи, которые в семантической структуре только лишь соединяют узлы, фактически никак не характеризуя их текстовую зависимость по смыслу.

Эти различия объясняются тем, что в зависимости от цели, для которой разрабатывался соответствующий метаязык, исследователь останавливался на том или ином уровне обобщения конкретных содержательных отношений, наблюдаемых между словами, синтаксически связанными в тексте.

**Синтаксический анализ текста в машинном перевода**

Прежде чем описывать семантические отношения и связи в машинном переводе, начнём с такого понятия, как *синтаксический анализ предложений* (по-английски *parsing*). Суть этого процесса состоит в построении графа, «каким-либо образом» отражающего структуру предложения. На сегодня не существует единственно принятой системы принципов, на которых строится граф. Даже в рамках одной концепции взгляды отдельных учёных на зависимости между словами могут различаться. В данное время существуют несколько методов синтаксического анализа и модернизируются в плот до семантического анализа (semantic parsing). Наверно, прежде всего надо разделить способы построения графа (обычно — дерева) на грамматике составляющих (phrase structure-based parsing) и грамматике зависимостей (dependency parsing).

Представители первой школы разделяют предложение на «составляющие», далее каждая составляющая разбивается на свои составляющие — и так до тех пор, пока не дойдём до слов. Эту идею хорошо иллюстрирует следующий рисунок:


Рис.2. Пример разбора предложения грамматикой "составляющие".

Представители второй школы соединяют зависящие друг от друга слова между собой непосредственно, без каких-либо вспомогательных узлов:

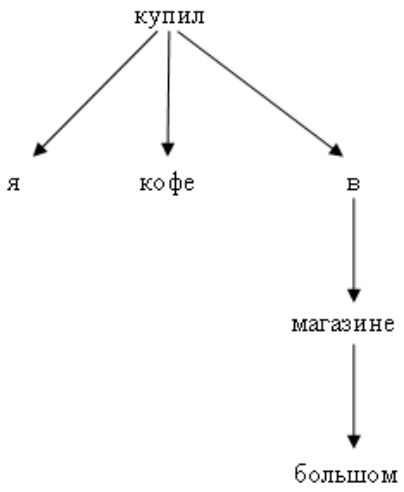


Рис. 3. Пример разбора предложения грамматикой "зависимости".

Вообще этот подход тоже трудно назвать особо свежим. Все ссылаются на работы Люсьена Теньера (Lucien Tesniere) пятидесятых годов как на первоисточник. Однако в компьютерной лингвистике dependency parsing долгое время был на втором плане, в то время как грамматики Хомского активно применялись. Вероятно, ограничения подхода Хомского особенно повлияли по языкам с более свободным (чем в английском) порядком слов, поэтому самые интересные работы в области dependency parsing до сих пор выполняются «снаружи» англоязычного мира. Так как русский язык относятся к группе со свободным порядком слов, разумно было применение для синтаксического анализа грамматики зависимости. и применение этого метода для казахского языка является удобным. [1,5]

Основная идея **dependency parsing** — соединять между собой зависимые слова. Центром практически любой фразы является глагол (явный или подразумеваемый). Далее от глагола (действия) можно задавать вопросы: кто делает, что делает, где делает и так далее. Основным качеством грамматики зависимости является, что соединяя между собой слова, не создаётся дополнительные сущности, и, стало быть, упрощается дальнейший анализ. В конце концов, синтаксический анализ — это лишь очередной этап обработки текста, и дальше надо представлять, что с полученным деревом делать. В каком-то смысле дерево зависимостей «чище», ибо показывает явные ***семантические связи*** между элементами предложения. Далее, нередко утверждают, что грамматики зависимости больше подходит для языков со свободным порядком слов. У Хомского все зависимые блоки так или иначе действительно оказываются рядом друг с другом. Здесь же в теории можно иметь связи между словами на разных концах предложения.

**Семантический анализ текста.**

Для семантического анализа текста будет использован предложенный метод расширенной атрибутной грамматики (РАГ), основанный на атрибутной грамматике Кнута [2]. С помощью грамматики зависимости будет построена дерево синтаксического анализа предложения. На входе семантико-синтаксического анализа текста в МП для дальнейшей работы необходимы начальные данные, полученные из лексико-семантического анализа методом РАГ. На данном этапе с элементов (слов) входного текста будут считываться полученные лексические свойства и семантические атрибуты для определения семантических отношении.

$$AAG = <G, A, R^W, R^F, R^S> \quad (1)$$

где G — контекстно-свободная грамматика предложений естественного языка, A – конечное множество семантических атрибутов; $R^W$– множество семантических правил на уровне слов, $R^F$- множество семантических правил на уровне фраз предложения, $R^S$- множество семантических правил на уровне предложения.

На основе метода РАГ были исследованы и предложены основные группы семантических атрибутов: действие ($A_{sem}$ (act)), субъект ($A_{sem}$ (sub)), объект ($A_{sem}$ (obj)), время($A_{sem}$ (tm)), места($A_{sem}$ (pl)), характеризирующие параметры ($A_{sem}$ (ch.pr)).

Можно сказать что выше предложенные семантические атрибуты пересекаются выше указанных научных работах. Но надо отметить что сам метод представления семантики и описания семантических отношении отличен от других. Предлагаемая расширенная атрибутная грамматика в отличии от других семантических методов рассматривает семантические свойства текста на различных уровнях анализа (лексический, синтаксический). Так как на разных этапах обработки текста свойственны различные типы семантических атрибутов и их связи, которые используются на следующем уровне анализа.

**Семантические отношения**

Для таких систем семантического описания, в которых значения слов представляются в виде элементарных смысловых единиц (сем), связанных элементарными семантическими отношениями естественно встает вопрос: как соотносятся между собой два множества — множество семантических отношений между значениями слов во фразе? На этот вопрос в принципе может быть два ответа:

1) СО между семами в описании значения слова и СО между значениями слов в тексте - это два разных множеств СО. В первом случае относится к тезариусу или системам толкования слов, то второе к семантическому анализу текста. Но надо учитывать что эти множества взаимосвязаны ;

2) все СО между значениями слов в тексте могут быть сведены к минимальному количеству СО.

Только второй ответ на этот вопрос соответствует задаче построения интегрального семантического описания языка. Такое описание предполагает, что фразы, признаваемые имеющими одно и то же значение, должны получить тождественные описания.

В методе РАГ составляется множество словосочетаний (фраз) $\{f_k\}$-, которые несут смысловые связи. Для определения семантических значений фраз и предложений вводятся семантические правила для группы существительного, глагола, обстоятельства, а также структуры предложений с учетом особенностей грамматики русского и казахского языка.

$$f_k=\{A_{sem}(w_i), A_{sem}(w_j)\} \quad (2)$$

На основе исследования **смыслового соединения фраз** были разработаны следующие основные структуры словосочетания, построенные на семантических правилах на уровне фраз $\boldsymbol{R^F(A)}$, с помощью которых воспроизводится простые семантические связи:

$$R^F(A)::= F \quad (3)$$

где F - множество семантических фраз

$$F:=\{f_k\},\ k=1,..,n. \quad (4)$$

F:={{$A_{sem}$ (ch.par),$A_{sem}$ (sub)} ,{ $A_{sem}$ (obj), $A_{sem}$ (act)} ,{ $A_{sem}$ (sub),$A_{sem}$ (pl)},{ $A_{sem}$ (tm),$A_{sem}$ (obj)}...}

Структура и связи базовых фраз были построены на основе грамматических правил русского и казахского языка с учетом смыслового соединения. Предложенным методом было разработано для казахского языка 26 , а для русского языка 36 смысловых структурных фраз

[2]. А так же в каждой фразе определяется доминирующее по смыслу слово, которое способствует в дальнейшем при построении семантической структуры метаязыка.

При переводе и генерации текста на казахский язык надо учитывать некоторые лингвистические свойства. Например: слова являющиеся семантическим атрибутом характеристического параметра ($A_{sem}$ (ch.par)) всегда расположен перед описываемым объектом (субъекта, места, действия, времени). Для глаголов или слов описывающие семантический атрибут действия ($A_{sem}$ (act)) по структуре всегда следует после других частей речи и семантических атрибутов. Связь проверяется с право на лево от основных смысловых элементов, т.е. от слова у которого семантический атрибут действия ($A_{sem}$ (act)).

Учитывая грамматические правила русского и казахского языка и смысловые взаимосвязи семантических атрибутов можно сказать, что полната базовых фразовых структур является достаточной. Основой этих отношений выступает ***дистрибуция*** ( дистрибутивный анализ). Связь между словами определяется по их расположению в речи относительно друг друга (сочетаемость, аранжировка) . Они формализуются с помощью математической теории вероятностей, статистико-вероятностного подхода, исчисления предикатов и исчисления высказываний, теории алгоритмов. Конечно можно было использовать выше перечисленные методы, но на качестве контекста предложения и определении семантических связей скажется плохой результат. Так как для естественных языков не существует единого математического аппарата описывающий все возможные вариации представления текста (синтаксический и семантический анализ).

При семантико-синтаксическом анализе текста мы получаем некое множество все возможных сочетаний семантических фраз предложения, которые дают основные связи между элементами на метаязыке для формирования смысловой онтологии текста. К множеству F будет применены ***семантические правила на уровне предложения*** $\boldsymbol{R^S}$**,** основанные на СО.

$$R^S(F) := O(S) \qquad (5)$$

где O(S)-онтология предложения S.

Таблица 1. Виды семантических отношении метода РАГ

| **Название** | **Структура** | **Пример** |
|---|---|---|
| Действие (action) | <x, $A_{sem}$ (act)> | Машина уехала -< машина, уехала ><br>Девочка устала - < девочка, устала> |
| Принадлежность (belonging) | <x, y> | Роман Абая -< Абая, роман ><br>Жители поселка -< поселка, жители > |
| Время (time) | < $A_{sem}$ (tm), $A_{sem}$ (act)> | Это произошло вчера -<br>< вчера, произошло > |
| Описание \значение (Specification) | < $A_{sem}$ (ch.p), y> | Красивое платье- < красивое, платье ><br>Быстро приехал -< быстро, приехал > |
| Инструмент \ средство (instrument\ means ) | <x, $A_{sem}$ (act)> | Резать ножом -< ножом, резать ><br>Гордиться страной - <страной, гордиться > |
| Расположение (location) | <x, $A_{sem}$ (pl)> ,<br>< $A_{sem}$ (pl), $A_{sem}$ (act)> | Яблоки из Алматы - < яблоки, Алматы ><br>Отдыхать на море- < море , отдыхать > |
| Имя (name) | <$A_{sem}$ (sub), $A_{sem}$ (sub)> | Дворник Степанов -< Степанов, дворник ><br>Я учитель -< я , учитель > |
| Причина \ цель (reason\ purpose) | <x,y> | Карантин в целях профилактики-<br>< карантин , профилактика ><br>Самолет не вылетел из за тумана -<br>< туман, не вылетел > |

В таблице 1 показаны основные виды СО, построенные на связях и семантических атрибутах элементарных смысловых единиц текста.

**Алгоритм построения семантических отношений**

Основой для алгоритма определения и построения СО будит служит грамматика зависимости, так как синтаксический анализ текста строится на этом методе. Основным (ключевым) объектом текста будут слова описывающие действие (глагол) и имеющие семантический атрибут $A_{sem}$(act). Надо учитывать что над текстом произведен лексико-семантический анализ, были определены и присвоены семантические атрибуты к словам предложения. Алгоритм определения СО показан на рисунке 4.

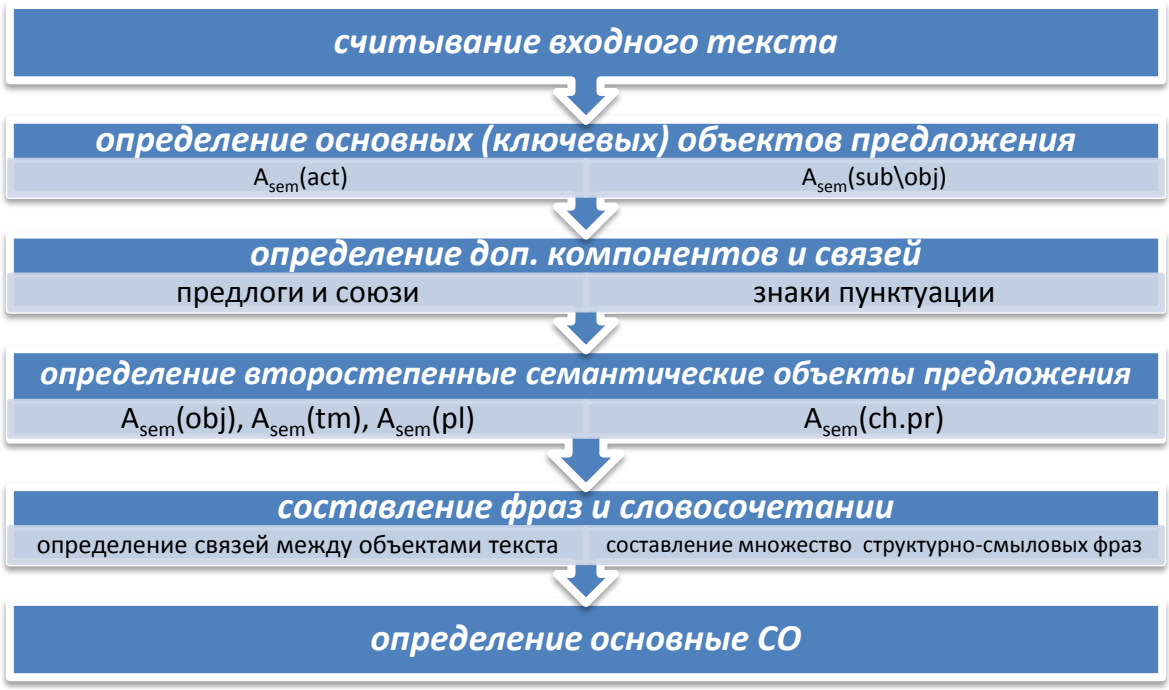


Рис.4. Алгоритм определения СО.

На рисунке проиллюстрирована общая схема принципа определения СО в тексте (для простых предложении). Конечно, каждый процесс является сложной системой со своими условиями и правилами.

Рассмотрим каждый модуль алгоритма:

1.***Считывание входного текста***. Определение количества объектов в тексте и длину объектов. Считывание синтаксические и семантические атрибуты объектов текста.

2.***Определение основных (ключевых) объектов предложения***. В тексте определяются слова имеющие семантический атрибут действия $A_{sem}$(act). И от этого ключевого объекта в лево производится поиск слова с семантическими атрибутами субъекта или объекта. В ином случайе поиск буден произведен в право от ключевого слова. После нахождения подходящего слова и ему присваивается статус ключевого слова в предложении. Потому что все связи и СО будут строится от этих компонентов. Конечно, доминирующим объектом в тексте будит слова с семантическим атрибутом действия. В случае отсутствия глагола и слов действия поиск и связи СО будут определятся от ключевых слов с атрибутами $A_{sem}$(sub\obj).

3.***Определение дополнительных компонентов и связей***. На смысл и структуру предложения текста так же очень может влияет предлоги и союзы , а так же пунктуации. В этом модуле описывается правила соединения и перевода на выходной язык.

4.***Определение второстепенные семантические объекты предложения***. После определения ключевых семантических объектов и общей структуры предложения надо

определит остальные объекты текста. В данном модуле применяется определенный набор правил и исключений для определения объектов с семантическими атрибутами $A_{sem}$(obj), $A_{sem}$(tm), $A_{sem}$(pl), $A_{sem}$(ch.pr) .

5.***Составление фраз и словосочетании***. В данном модуле определяются связи между объектами и создаются все возможные структурно -смысловые фразы предложения.

6.***Определение основные СО***. Из всего множества фраз надо определить основные фразы несущие основной контекст предложения и довести до минимума количество СО.

При синтаксическом и семантическом анализе предложения (текста) выявляется множество фраз и словосочетании, конечно ни все эти соединения несут смысловое значение**.** И связи с этим надо оптимизировать множество фраз (F). А так же все СО между значениями слов в тексте могут быть сведены к минимальному количеству СО.

Введем обозначение F*- является множеством фраз , элементы которых пересекаются хотя бы один раз.

$$F^*=f_i \cap f_j$$

Таким образом можно определить семантические узлы, с помощью которых можно вычислить связи и семантические отношения. Эффективность построения СО и скорость поиска на много увеличивается.

Пример:

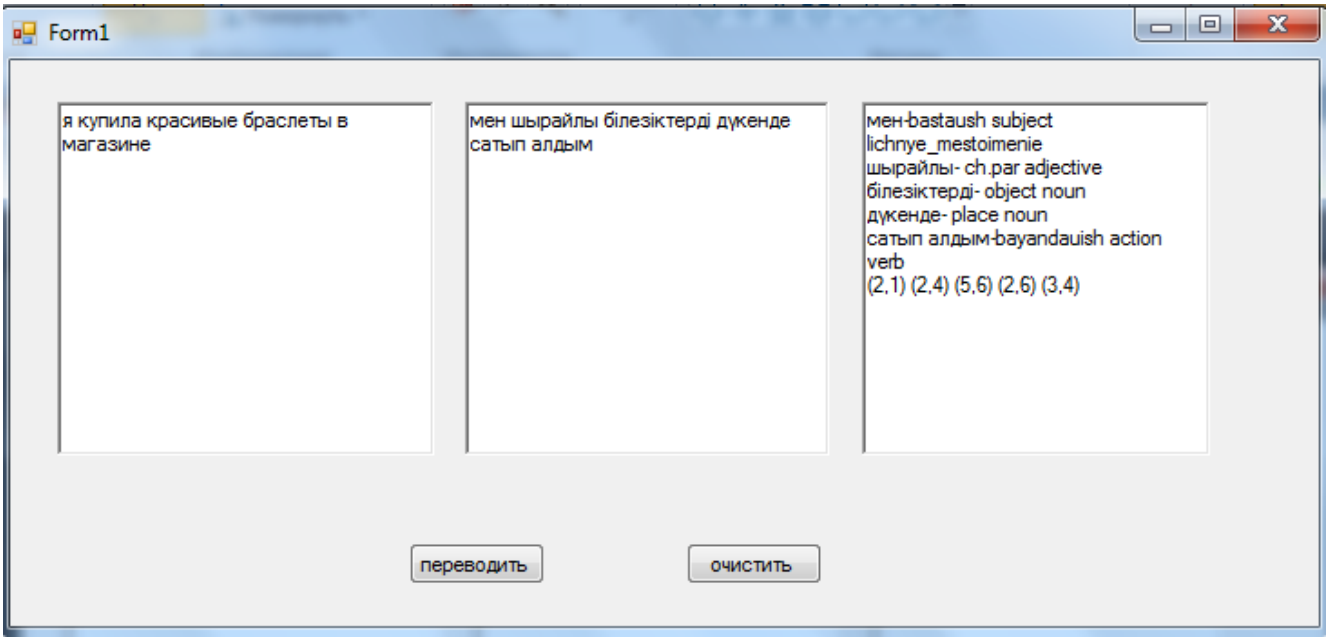


Рис.5. Пример машинного перевода и семантического анализа простого предложения с русского на казахский язык.

На рисунке показан входной текст и полученный результат на выходном языке. В третьем окне показан семантический анализ предложения и составленные СО. СО отмечены порядковыми номерами объектов (единиц) входного текста (русского языка).

**Заключение**

Основной семантической задачей машинного перевода является полный смысловой анализ текста на метаязык с помощью которого будут сгенерированы предложения на выходной язык. При обработке текста в машинном переводе на начальных и на отдельных стадиях были свои проблемы и затруднения, которые можно было решить с помощью дополнительных семантических методов. В данной работе были рассмотрены типы и способы

представления СО. Проведен анализ научных работ по СО. А так же в данной работе был представлен метод семантического анализа (РАГ), с помощью которого определяются семантические атрибуты объектов текста , связи и СО между ними. Разработан алгоритм представления СО при семантическом анализе текста. Приводится пример практического использования предлагаемого метода.

## Список литература


1.Тукеев У.А., Рахимова Д.Р. Синтаксический анализ казахского языка на основе грамматики зависимости. Труды международной научно-практической конференций. «Информационные и телекоммуникационные технологии: образование, наука, практика», КазНТУ имени К.И. Сатпаева 2012 г.

2.Tukeyev U., Rakhimova D. "Augmented attribute grammar in meaning of natural languages sentences", SCIS-ISIS 2012 The 6th International Conference on Soft Computing and Intelligent Systems. The 13th International Symposium on Advanced Intelligent Systems, pp 1080-1084.

3.Семантическая сеть. http://ru.wikipedia.org/ (обращение10.08.2013)

4. "Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ)" http://www.aot.ru/docs/SemRels.htm (обращение 10.08.2013)

5.Заметки об NLP . http://habrahabr.ru/post/79830/ (обращение 20.08.2013)

6. Найханаова Л.В. Основные типы семантических отношений между терминами предметной области, Известия высших учебных заведений. Поволжский регион. Технические науки. №1 2008.

7.www.durov.com/study (обращение 05.09.2013)