**Математическая модель естественного языка в задачах нечеткого   
ассоциативного поиска**

*Крайванова В.А., Кротова А.О., Крючкова Е.Н.*

Алтайский государственный технический университет им. И.И. Ползунова

**Abstract**

Рассматриваются проблемы ассоциативного поиска. Знания системы основаны на толковом словаре и словаре синонимов, которые используются для автоматической обработки текстов.

Problems of associative search are considered. The system knowledge are based on the explanatory and synonym dictionaries which are used for automatic text processing.

**Введение**

Рассмотрим вид информационного поиска в Интернет, основной целью которого является совершение определенного выбора на основе некоторых приблизительно сформулированных требований к предмету поиска. Примерами таких задач являются поиск подарка на определенный праздник, поиск товара в интернет-магазинах, поиск нового сотрудника. Эта разновидность поиска обладает следующими особенностями:

* у пользователя имеются некоторое представление о предмете поиска, однако свои требования он не может точно описать;
* на основе имеющегося опыта пользователь может иметь ассоциативные представления о предмете поиска, которые достаточно сложно сформулировать в форме поискового запроса, например, в Google ;
* успешное решение приведенных задач требует от пользователя достаточно полного представления об информационной структуре предметной области;
* каждый раз при обращении к этим задачам пользователь может обновлять свои знания о предметной области.

Получать информацию из текстов, размещенных в Интернет, для данной проблемы очень трудоемко. Система поиска, основанная на связях между понятиями предметной области и их визуализации, может существенно ускорить и облегчить выполнение описанных задач.

Сформулируем требования к модели знаний:

* Модель должна быть достаточно универсальной относительно предметной области.
* Модель должна обладать сравнительно простой структурой (что позволит составить алгоритмы автоматизированного извлечения знаний из текстов на естественном языке).
* Должны существовать алгоритмы автоматического построения физического представления модели.
* Модель должна быть ориентирована на построение по достаточно большому корпусу текстов (чтобы обеспечить достоверность содержащихся в ней знаний)
* Модель должна иметь возможности репрезентативной визуализации.

Основой предлагаемой модели является лексикон. Рассматриваемая структура лексикона представляет собой развитие модели, описанной в работах [2] и [3]. Источником лексикона могут служить существующие лингвистические словари. В данной работе рассматривается система, построенная на автоматической обработке толкового словаря Ожегова [4] и словаря синонимов [5]. Возможно использование любых других неадаптированных для автоматической обработки словарей.

**Структура лексикона**

Пусть *A* - алфавит лексического уровня применяемого естественного языка (ЕЯ). Тогда формально словом будем называть конечную упорядоченную последовательность непустых цепочек над этим алфавитом, разделенных пробелами. Пусть *W⊂A+* - лексикон модели. Слова в лексиконе *W* могут являться как отдельными словами ЕЯ, так и устойчивыми словосочетаниями. Абстрагируемся от морфологических особенностей словоформ, и будем считать, что слова ЕЯ приведены к словарной форме.

Строго говоря, лексикон ЕЯ - бесконечное, динамически изменяющееся множество, но множество слов, содержащихся в словарях, является конечным. Поэтому множество *W* также будем считать конечным.

На множестве *W* зададим три типа отношений, которые мы можем извлечь из словарей:

* отношение синонимии;
* отношение определения;
* отношение ассоциации.

Поскольку в рассматриваемых словарях определения и синонимы даются для слов естественного языка, а не для научных терминов, разумно считать их не абсолютно точными, а заданными с некоторой степенью неопределенности. Поэтому вводимые отношения будем считать нечеткими.

*Отношение синонимии.*

Слово *wB* является синонимом к слову *wA*, если для любой фразы *ϕ*, содержащей слово *wA*, фраза *ϕ'*, в которой на месте некоторых вхождений слова *wA* стоит слово *wB*, совпадает по смыслу с фразой *ϕ*. Существующие словари синонимов позволяют автоматически построить соответствующие отношения. Будем считать, что нечеткое **отношение синонимичности** *Syn: W×W → [0..1]* определяет степень уверенности системы в том, что некоторое слово *wB* *∈ W* является синонимом к слову *wA* *∈ W*.

*Отношение определения*

Слово *w* из фразы *ϕ* назовем **главным словом** этой фразы, если в фразе *ϕ* отсутствуют словосочетания, в которых слово *w* является зависимым (подчиненным). Если таких слов несколько, то главным будем считать первое по порядку следования слово. Слово *wA* является определением к слову *wB*, если слово *wA* является главным для фразы *ϕ*, определяющей слово *wB* в толковом словаре. Нечеткое **отношение определения («это есть»)** *Def: W×W → [0..1]* определяет степень уверенности системы в том, что некоторое слово *wA* *∈ W* является определением к слову *wB* *∈ W*.

В большинстве случаев отношение определения связывает слово с обобщающим его понятием (гиперонимом).

*Отношение ассоциации*

Слово *wA* назовем ассоциацией к слову *wB*, если слово *wA* присутствует в фразе *ϕ*, определяющей слово *wB* в толковом словаре, но при этом не является главным словом этой фразы. Нечеткое **отношение ассоциации** *Acс: W×W → [0..1]* определяет степень уверенности системы в том, что некоторое слово *wA* *∈ W* является ассоциацией к слову *wB* *∈ W*.

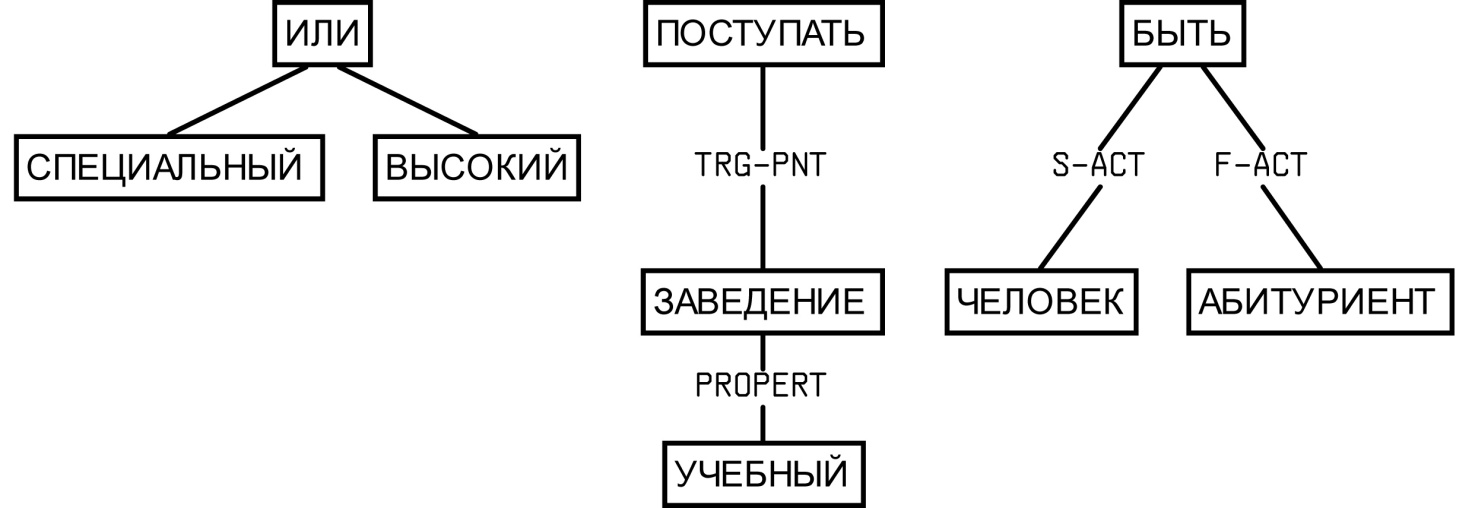
Введенные отношения позволяют установить степень близости слов на основе словарей. Назовем величину *Dict (wA, wB)* **словарным уровнем близости слов** *wA* и *wB. Dict (wA, wB) = Max (Syn (wA, wB), Def (wA, wB), Acc (wA, wB)).*

Выбор именно этих видов отношений мотивируется в первую очередь тем, что приведенные типы связи естественным образом извлекаются из словарей. Для разбора словаря синонимов и построения отношений *Syn* не требуется синтаксический анализ естественного языка. Значение *Syn (wA, wB)* зависит от количества синонимов слова *wA* и от удаленности синонима от определяемого слова в словарной статье. В отличие от словаря синонимов, статьи толкового словаря требуют синтаксического анализа, который облегчается тем, что определения в толковом словаре имеют очень строгую структуру. Эксперименты с анализатором Dialing [6] показали, что определения представляют собой два вида конструкций: для глаголов и для именных форм.

Если определяемое понятие представляет собой инфинитив, подробный анализ семантического графа не требуется. Отношение *Def* строится с инфинитивами, стоящими в определении. С остальными словами строится отношение *Acс*.

Если определяемое слово – именная форма, то наиболее часто встречаются определения, аналогичные по структуре следующей словарной статье:

*Абитуриент - человек, поступающий в высшее или специальное учебное заведение*.

Рисунок 1. Результат пресемантического разбора предложения программой «Dialing»

Связка отношений F-ACT(0,1) – S-ACT(X,1) однозначно приводит в предложениях подобной структуры к подлежащему (или запятой, союзу между ними) в узел *Х*. С этим словом (словами, соединенными запятой или союзом) стоится отношение *Def*. С остальными словами – *Acс*.

Сделанные предположения подтвердились на практике – потери от «нестандартной» структуры составили порядка 10% словарных статей, что является хорошим показателем.

Слова, связанные отношениями *Acc*, *Def* и *Syn*, образуют взвешенный граф. На основе трёх рассмотренных отношений возможно построение более строгих и семантически сложных отношений, таких как отношение обобщения, отношение часть-целое и др.

**Построение лексикона**

Реальный уровень близости слов существенно зависит от предметной области, в которой рассматриваются слова. Например, в компьютерной терминологии слово ВИРУС гораздо теснее связано со словом ПРОГРАММА, чем в биологии. Скорректировать уровень близости слов для каждой предметной области предлагается через анализ текстов данной области.

**Синтаксическим расстоянием** *SD(wA, wB)* двух слов *wA* и *wB* в предложении *ϕ* назовем длину пути между этими словами в дереве синтаксического разбора предложения *ϕ*.

Определим понятие **синтаксической близости** *Prox(wA, wB)* двух слов для коллекции текстов. Пусть *Sent(w1, w2,…wn)* – множество предложений, одновременно содержащих слова *w1, w2,…wn*. Тогда смысловой близостью назовем величину

На основе синтаксической близости корректируется уровень словарной близости для отдельных предметных областей.

**Ассоциативные окрестности слов**

Очевидно, что чем дальше друг от друга в лексиконе находятся слова, тем менее они связаны. Пусть *L(x,y) -* расстояние в графе между понятиями *x и y,* которое определим как максимальное значение произведения весов на пути между *x и y*. Семантической α-окрестностью слова *x* в лексиконе *W* назовем множество слов, которые находятся от него на расстоянии, не превышающем α:

*D(x, α) = {y | y ∈ W , L(x,y) < α }.*

Выделение некоторого понятия в α-окрестности позволяет построить новую α-окрестность, выполняя ассоциативный поиск и ассоциативный логический вывод в рассматриваемой предметной области.

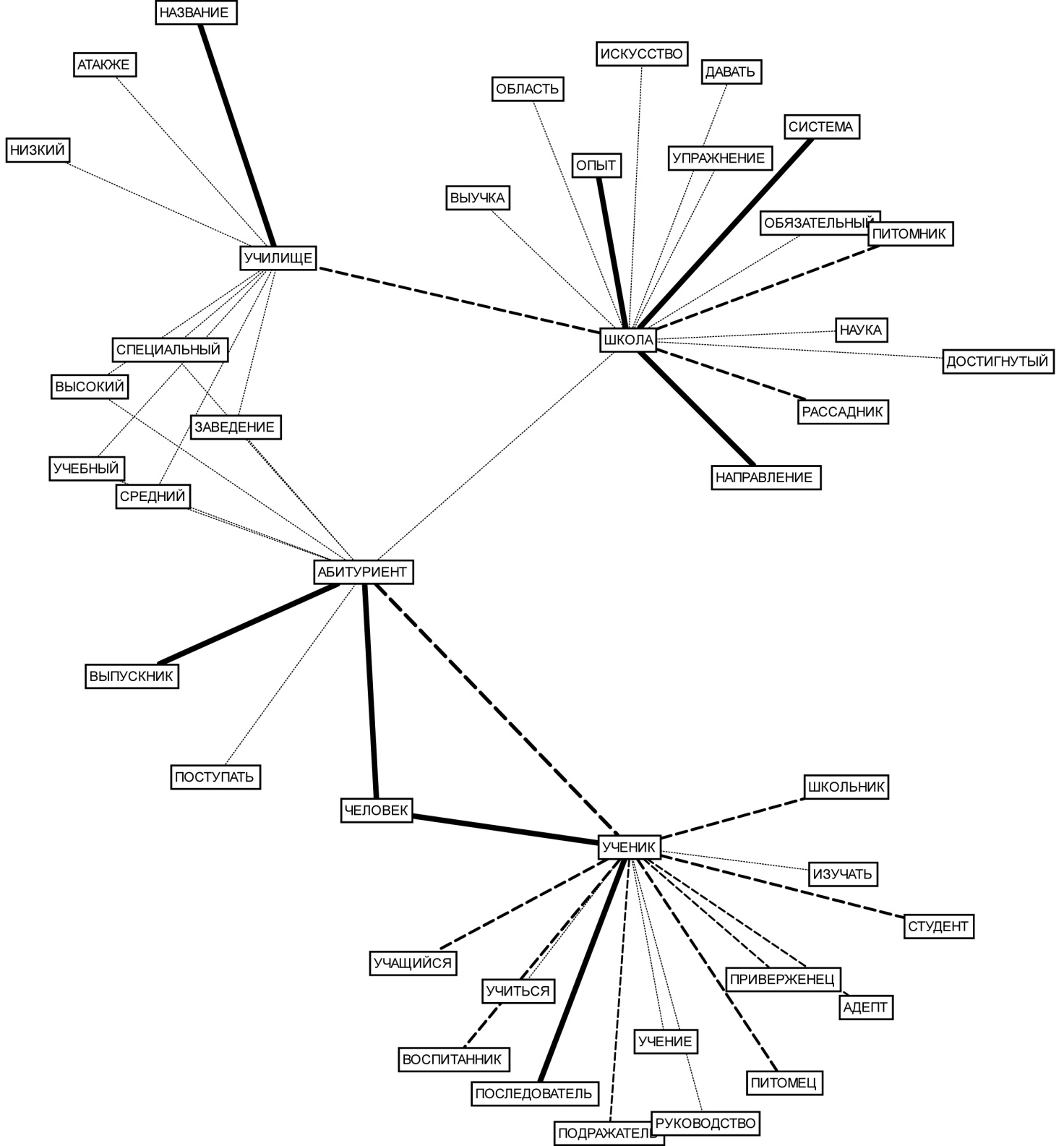


Рисунок 2. Часть графа вокруг слова «Абитуриент»

На рисунке 2 представлена часть словарного графа. Сплошными линиями обозначены связи *Def*, пунктиром – связи *Syn*, точками – связи *Acc*. Толщина линии соответствует весу связи. Для построения графов использовалась программа AiSee[7].

**Литература**

1. 4 и, по-хорошему, что-нибудь еще из этой области, например, как ссылка к первой части, где описывается проблема
2. Krayvanova V., Kryuchkova E. The mathematical model of the semantic analysis of phrases based on the trivial logic // In Proceedings of the 13-th International Conference ''Speech and Computer'' SPECOM'2009, pp. 543-546
3. Крайванова В.А., Крючкова Е.Н. Проблема извлечения знаний в условиях динамически изменяющейся внешней среды на примере модели логического анализа текстов на естественном языке // Вычислительные технологии. 2010. Т. 15. № 3. С. 99-108
4. Ожегов С.И., Шведова Н.Ю. Толковый словарь русского языка [Электронный ресурс]. Издательство "Азъ", 1992 – Режим доступа: http://lib.ru/DIC/OZHEGOW/
5. [Абрамов](http://www.ozon.ru/context/detail/id/2889654/#persons) Н. Словарь русских синонимов и сходных по смыслу выражений [Электронный ресурс]. Издательство [Русские словари](http://www.ozon.ru/context/detail/id/858227/), 2007 г – режим доступа: http://dict.buktopuha.net/data/abr1w.zip
6. Сайт рабочей группы «Автоматическая обработка текстов» [Электронный ресурс] – Режим доступа: http://aot.ru
7. Программа визуализации графов AiSee [Электронный ресурс] – Режим доступа: http://www.absint.com/aisee/index\_ru.htm