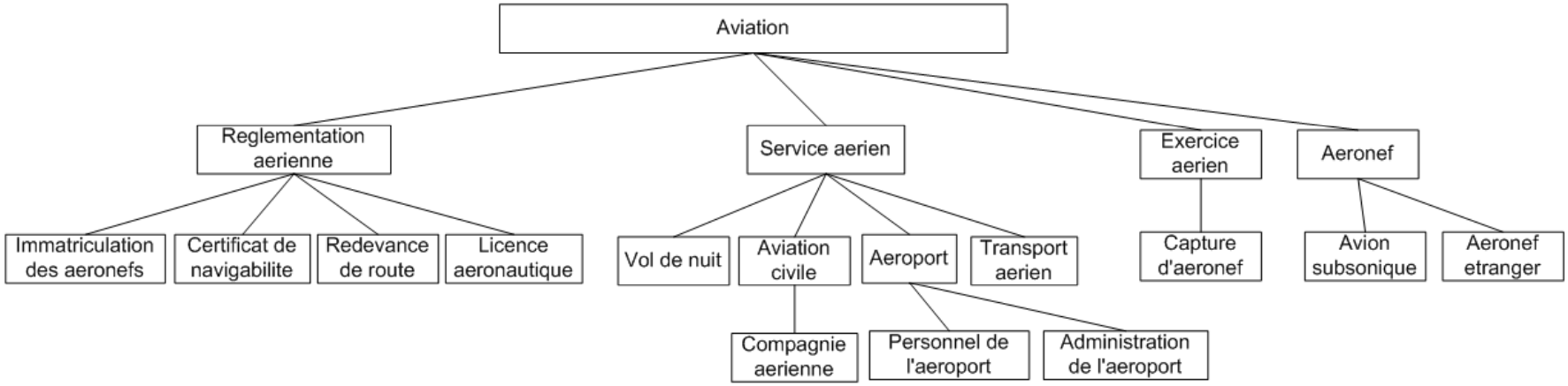
Построение семантической сети из коллекции текстовых документов

Аспирант: Александр Панченко

Научный руководитель:
к.т.н. Юрий Николаевич Филиппович

Основные понятия

- **Семантическая сеть** —модель предметной области, имеющая вид ориентированного графа, вершины которого соответствуют объектам предметной области, а дуги задают отношения между ними.
- **Тезаурус**— словарь общей или специальной лексики, в котором заданы семантические отношения (синонимы, гипонимы, гиперонимы, ассоциативные связи) между лексическими единицами



Цели работы

- Разработка **технологии** построения семантической сети из коллекции текстовых документов.
- Разработка и применение **критериев качества** построенной сети на основе сравнения с **тезаурусом**, составленным вручную.
- Реализация **программного средства** на основе предлагаемой технологии.

Актуальность работы

Технологию построения семантической сети можно применить для

- **Автоматизации** создания семантических ресурсов, таких как информационно-поисковый тезаурус.
- Улучшения **качества информационного поиска** в корпоративных и специализированных поисковых системах (рис.1).
- Для **визуализации** и улучшенной **навигации** по коллекции текстовых документов (рис.2)*.

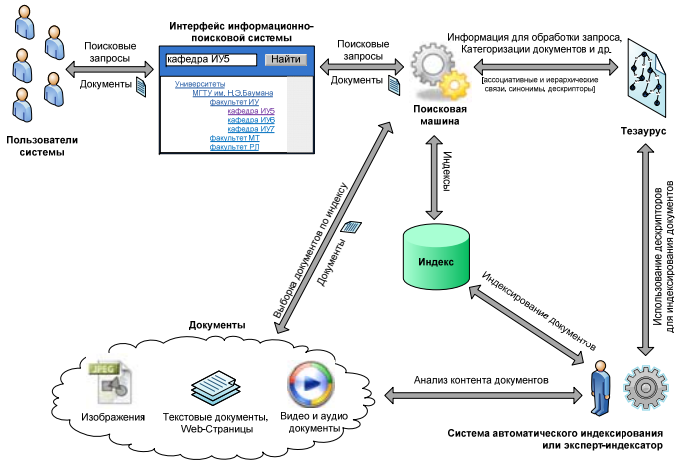


Рис.1

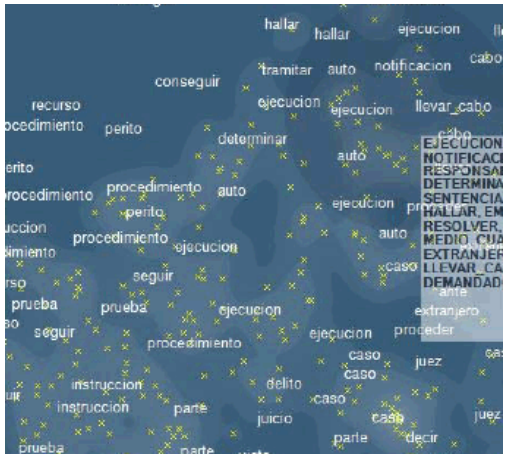


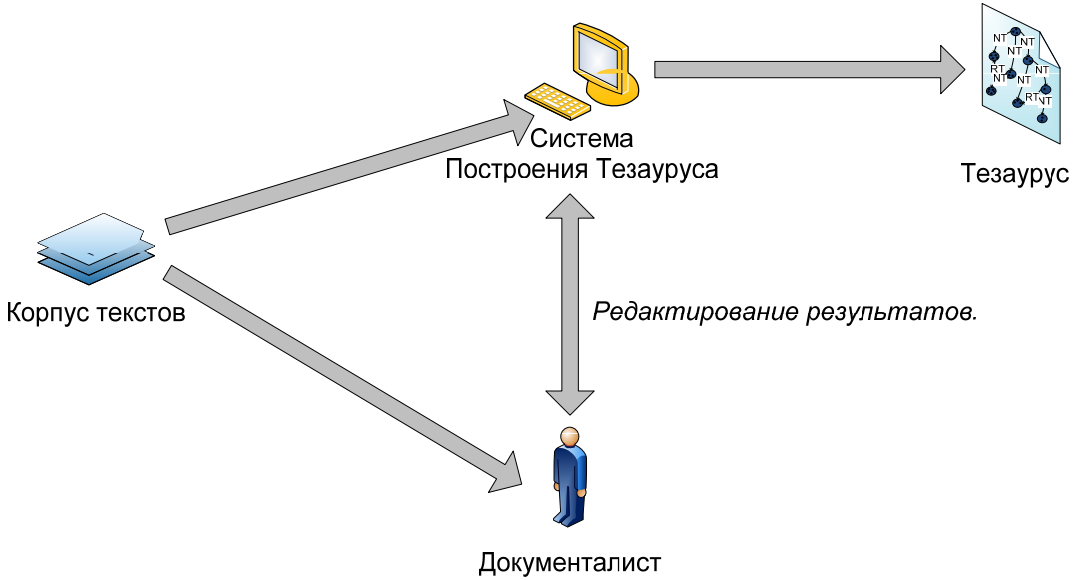
Рис.2

*Иллюстрация из B. Fortuna, M. Grobelnik and D.Mladenić, Visualization of Text Document Corpus , Informatica 29 (2005) 497–502

Автоматизация создания семантического ресурса

Ручное составление тезауруса

Автоматизированное составление тезауруса



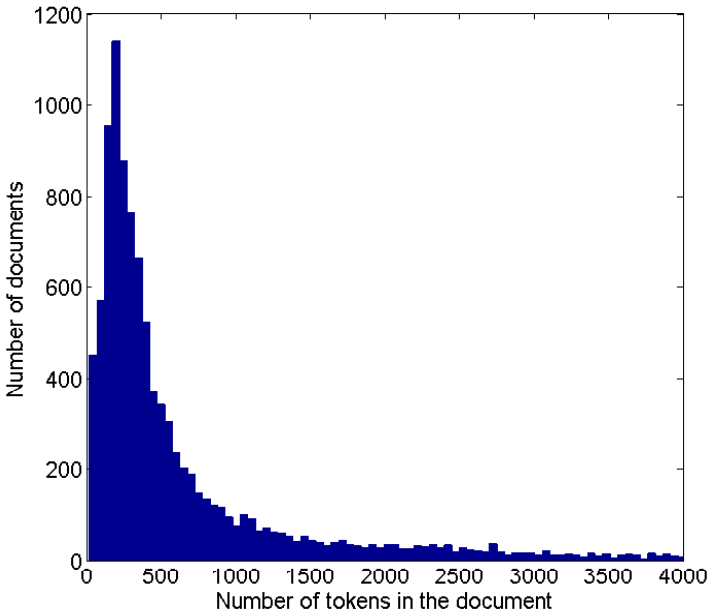
Набор данных

1. **Корпус** политических текстов на французском языке: депутатские запросы, протоколы заседаний парламента и т.п.

   11.382 тестовых документа

   20.665.146 токенов (wc)

   длина документа в среднем 1-3 стр.



2. **Тезаурус**, составленный вручную, описывающий предметную область.

   2.514 концептов $C = \{c_1,..,c_k\}$ каждый из которых состоит из мн-ва дескрипторов

   $c_i = \{d_{i1},..,d_{i8}\} = \{'Самолет',...,'Дирижабль'\}$

   4.771 дескрипторов $r_i = \langle d_1, d_2 \rangle, d_i \in D$

   2.456 иерархических и 765 ассоциативных отношений

Научная новизна

- **Оценка качества** построенной семантической сети на основе сравнения с семантическим ресурсом, построенным вручную.
- Исследование эффективности применения методов **машинного обучения** для вычисления метрик адоптированных для выявления семантической связи **определенного типа**.
- Качественное извлечение не только ключевых слов, но и ключевых словосочетаний
- Исследование эффективности применения моделей **тензорного счисления** для выявления семантических связей между ключевыми понятиями предметной области.

Технология построения семантической сети

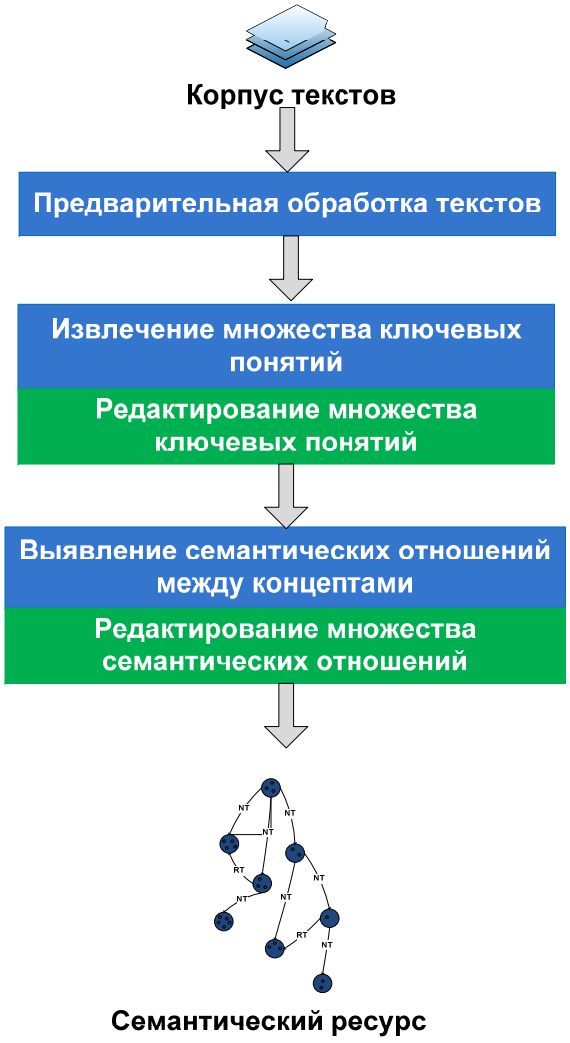
1. Предварительная обработка корпуса текстов

2. Извлечение множества **ключевых понятий** предметной области (концептов) из корпуса текстов
   - Выделение ключевых слов
   - Выделение ключевых словосочетаний

3. Выявление семантических **отношений** между концептами

Технология построения семантической сети

Предполагает возможность **контроля** документалистом, составляющим семантический ресурс



Автоматическая обработка

Ручная обработка

Технология: предварительная обработка корпуса текстов (1)

Цель – трансформировать свободный текст в представление, пригодное для дальнейшей алгоритмической обработки.

Средства – программные средства обработки естественного языка: синтаксические анализатор, лемматизатор и т.п.

В нашем исследовании используются Unitex[1], Xerox Incremental Parser[2].

Состав определяется потребностями применяемых моделей и алгоритмов.

[1]http://www-igm.univ-mlv.fr/~unitex/

[2]http://www.xeroxtechnology.com/ip1.nsf/sedan1?readform&unid=9C7EE64CFD78931585256FCD005C454D&nav=nav_cat_7

Технология: предварительная обработка корпуса текстов (2)

1. **Нормализация** текстовых документов: приведение к общему формату, кодировке, декапитализация, деакцентизация, очистка от метаданных и разметки.

2. **Токенизация** – установление границ слов и предложений.

3. **Лемматизация** – определение канонической формы слов.

4. Выделение **словосочетаний**.

5. Удаление **стоп-слов** и символов.

6. **Синтаксический анализ** текста.

Технология: выделение ключевых слов (1)

1. Нормализация, токенизация, лемматизация.

   ~149.500 лемматизированных слов

2. Фильтрация на основе **лингвистической информации**:
   - удаление стоп слов и служебных символов
   - удаление имен собственных
   - удаление чисел
   - удаление дат
   - оставляем только существительные и прилагательные

   ~78.250 лемматизированных слов (-50%)

Получаем множество **слов-кандидатов** $D_j = \{d_{1j}, d_{2j}, \ldots\}$ для каждого из документов $doc_j \in DOC$.

Множество всех слов-кандидатов для всей коллекции документов : $D = \bigcup_{j=1}^{|DOC|} D_j$

Прим.: *doc*= “В 2010 году "Булава" полетит с нового подводного крейсера.”
D= “булава полететь подводный крейсер”

Технология: выделение ключевых слов (2)

3. Ранжирование слов-кандидатов с использованием **статистической информации**:

Функция ранжирования $\Re : D \to \mathbb{N}$

А) По **частоте** $rank_i = n_i$, ключевые слова – первые *x%*

Б) **“Глобальный” TF-IDF**, ключевые слова – первые *x%*

$$rank_i = \frac{n_i}{\sum_{j=0}^{N} n_j} \log\left(\frac{|DOC|}{|\{doc : d_i \in doc\}|} \right), \qquad D_{key} = \left\{ d_i : \frac{x}{100}|D| \le rank_i \right\}$$

В) **“Локальный” TF-IDF**, ключевые слова – объединение первых *x%*, но не более чем *y* слов*

$$rank_{ij} = \frac{n_{ij}}{\sum_{j=0}^{N} n_{ij}} \log\left(\frac{|DOC|}{|\{doc : d_i \in doc\}|} \right), \qquad D_{key} = \bigcup_{j=1}^{|DOC|} \left\{ d_i : \min\left\{ y, \frac{x}{100}|D_j| \right\} \right) \le rank_{ij} \right\}$$

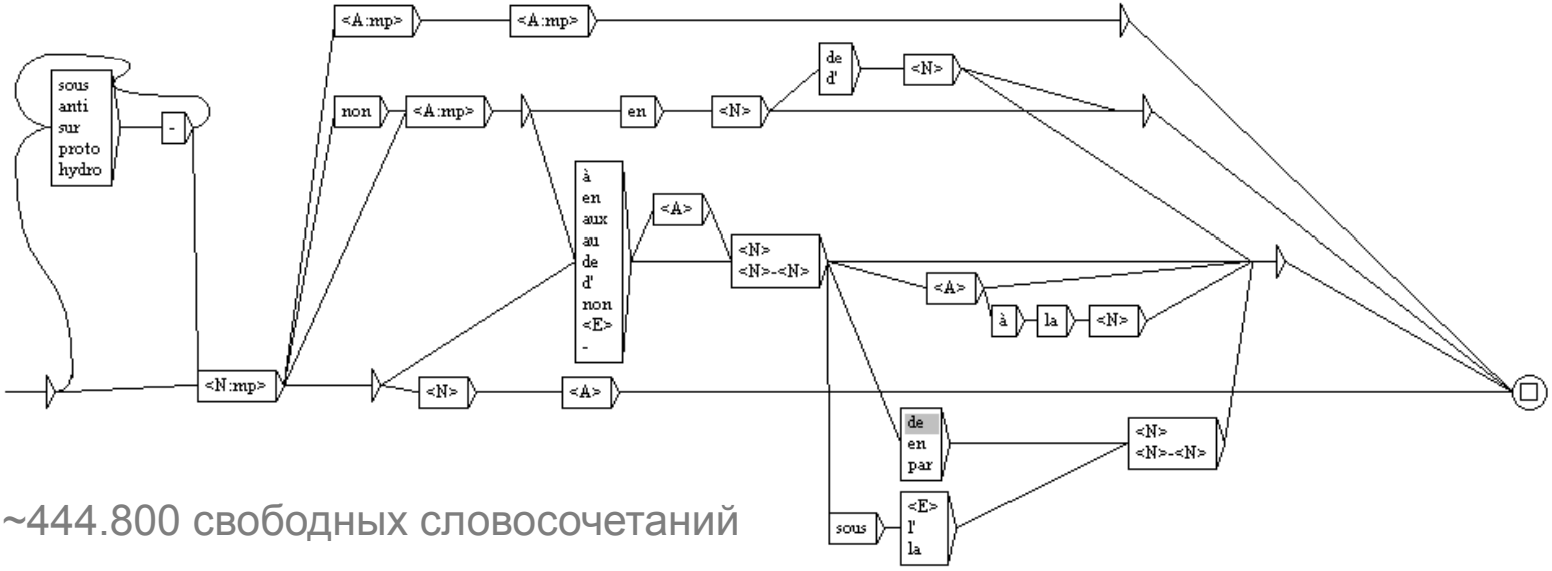
В наших экспериментах мы использовали с x = 33%, y = 50 слов

Технология: выделение ключевых словосочетаний (1)

1. Извлечение **свободных словосочетаний**

   Прим.: “**Прозрачный воздух** был теплым и нежным”

   Использование **символьного метода**: набор конечных автоматов, фиксирующих лингвистический феномен и словарей (Unitex).



~444.800 свободных словосочетаний

Технология: выделение ключевых словосочетаний (2)

2. Группирование **словосочетаний-кандидатов** путем поиска наибольших общих подстрок.

   Для каждого строки (словосочетания) d_i найти все строки, которые содержали бы содержали d_i : $\{d : d_i \subseteq d\}$

2. Ранжирование **словосочетаний** с помощью следующей формулы:

$$rank_i = g_i \frac{n_i}{\sum_{j=0}^{N} n_j} \log\left(\frac{|DOC|}{|\{doc : d_i \in doc\}|}\right),$$

   g_i – коэффициент группирования

ANCIEN CHEF
- ancien chef
- ancien chef de la centrale
- ancien chef d'etat
- ancien chef d'exploitation,
- anciens chefs de gouvernement
- ancien chef du laboratoire
- ancien chef politique
- ancien chef du service
- ancien chef des services secrets

3. Ключевые словосочетания – первые x%

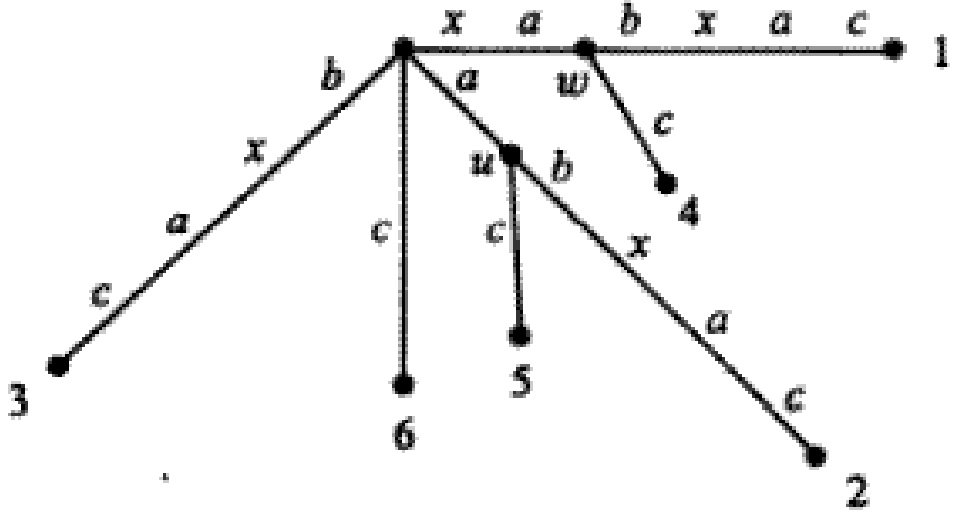
Технология: выделение ключевых словосочетаний (3)

Алгоритмическая сложность решения задачи (2) “в лоб” $O(n^3)$

Суффиксное дерево позволяет решить задачу (2) за время $O(n)$

Мы используем алгоритм Укконена[2] для построения дерева.

Суффиксное дерево *T* для *m*-символьной строки *S* – ориентированное дерево с корнем имеющее ровно m листьев. Для каждого листа *i* конкатенация меток дуг на пути от корня к листу *i* в точности составляет суффикс *S*, который начинается с позиции *i*.



Суффиксное дерево для строки “xabxac” *

*Иллюстрация из D. Gusfield, Algorithms on strings, trees, and sequences : computer science and computational biology, 2007.

Оценка качества ключевых слов и словосочетаний

Используем множество слов D_{golden} из составленного вручную тезауруса для оценки качества извлеченных слов и словосочетаний по критериям [4]

$$Точность = \frac{\left|\left\{D_{golden}\right\} \cap \left\{D_{key}\right\}\right|}{\left|\left\{D_{key}\right\}\right|}, \qquad Полнота = \frac{\left|\left\{D_{golden}\right\} \cap \left\{D_{key}\right\}\right|}{\left|\left\{D_{golden}\right\}\right|}$$

Экспериментальные результаты качества **ключевых слов**:

Метод ранжирования	**Нет**	**По частоте**			**“Глобальный” TF-IDF**			**“Локальный” TF-IDF**		
x	100%	10%	15%	33%	10%	15%	33%	10%	15%	33%
Полнота, %	92%	62	74	87	62	73	87	24	24	24
Точность, %	2%	13	10	5	13	10	5	17	17	17
Количество ключевых слов	73.513	7.351	11.027	24.259	7.351	11.027	24.259	1.196	2.198	2.198

$$\left|D_{golden}\right| = 1590$$

Технология: Выявление семантических отношений между концептами (1)

Задача: найти множество бинарных отношений между концептами

$$\hat{R} = \{r_1,..,r_n\}, r_i = \langle d_1, d_2 \rangle, d_i \in D_{key}$$

Множество отношений между ключевыми понятиями строиться с помощью метода **дистрибутивно-статистического анализа**, основанного на анализе совместной встречаемости **синтаксических контекстов**, в которых встречается то или иное ключевое понятие.

Метод состоит из следующих этапов:

1. **Предварительная обработка** тезауруса и корпуса текстов
2. **Индексирование** ключевых понятий.

   Каждый элемент индекса *I* содержит информацию об одном вхождении дескриптора d_i в документ δ_j, на позиции *b*

   $$\langle \delta_i, b, e, d_j \rangle, \delta_i \in DOC, \{b, e\} \in \mathbb{N}, d_j \in D_{key}$$

3. Извлечение множества **синтаксических зависимостей**

   $$SR_s = \{sr_1, sr_2, ...\}, sr_i = \langle w_i, b_i, t_j, w_k, b_k \rangle = \langle \text{Conseil}, 4, DETERM, \text{Le}, 0 \rangle$$

Технология: Выявление семантических отношений между концептами (2)

5. Вычисление мультимножества **синтаксических контекстов** *C*

$c_{ij} = \langle d_i, \beta_j \rangle$, где β_j – синтаксический признак, к примеру $\langle SUBJ, \text{proposition} \rangle$

и матрицы свойств **F** и матрицы подобия **S**

$$\mathbf{f}_i = \sum_{j=1}^{n} m(\langle d_i, \beta_j \rangle)\mathbf{b}_j, \qquad sim(d_i, d_j) = s_{ij} = \frac{\mathbf{f}_i \cdot \mathbf{f}_j}{|\mathbf{f}_i| \cdot |\mathbf{f}_j|}$$

6. Вычисление **множества отношений** между дескрипторами с помощью задания определенного порога для матрицы подобия.

$$\hat{R} = \{\langle d_i, d_j \rangle : s_{ij} \geq s_t\}$$

Технология: Выявление семантических отношений между концептами (3)

Алгоритм вычисления мультимножества синтаксических контекстов *C*.

Require: $SR = \bigcup_{s=1}^{N} SR_s$ – set of syntactic tuples, where SR_s is set of syntactic tuples for the sentence s,
I – vocabulary index,
$StopPOS$ – list of stop POS such as determiner, pronoun etc.
$StopWORD$ – list of stopwords.

$C \leftarrow \emptyset$
$w_{context} \leftarrow$ ""
$w_{descr} \leftarrow$ ""
for $s = 1$ to N **do**
  for all $\langle w_i, b_i, t_j, w_k, b_k \rangle \in SR_s$ **do**
    if $\exists \langle \delta, b, e, d \rangle \in I : b_i \in [b; e]$ **then**
      $w_{context} \leftarrow w_i$
      $w_{desc} \leftarrow d$
    else if $\exists \langle \delta, b, e, d \rangle \in I : b_k \in [b; e]$ **then**
      $w_{context} \leftarrow w_k$
      $w_{desc} \leftarrow d$
    else
      continue
    end if
    if $w_{context} \notin w_{desc}$ **and** $w_{desc} \notin w_{context}$ **and** $POS(w_{context}) \notin StopPOS$ **and** $w_{context} \notin StopWORD$ **then**
      $\beta \leftarrow \langle t_j, w_{context} \rangle$
      $C \leftarrow C \cup \leftarrow \langle w_{desc}, \beta \rangle$
    end if
  end for
end for
return C

Оценка качества множества найденных отношений между концептами

Используем три критерия:

точное совпадение, приблизительное совпадение и нечеткое совпадение

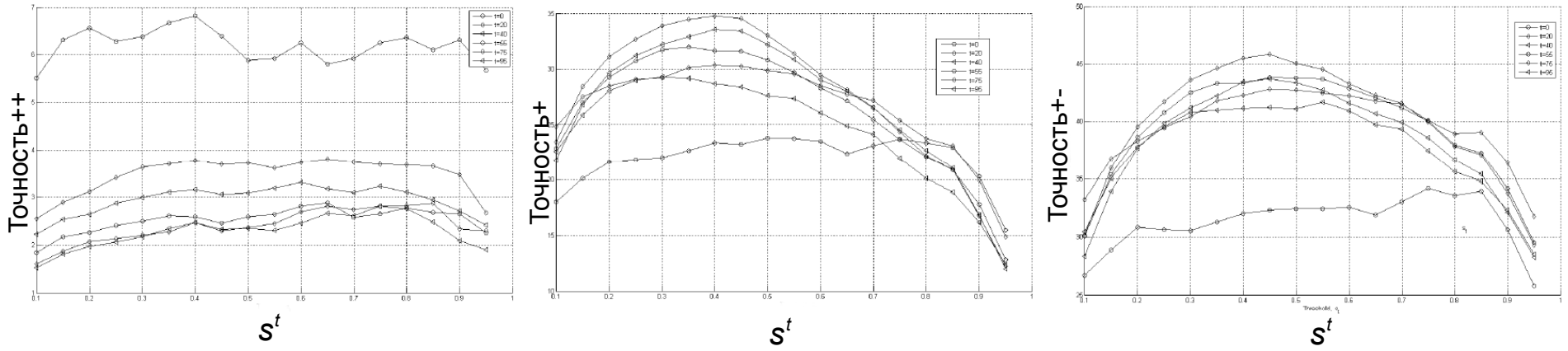
$$\textit{Точность}^{++} = \frac{|\hat{R} \cap R|}{|\hat{R}|}, \quad \textit{Точность}^{+} = \frac{|\hat{R} \cap R_{fuzzy1}|}{|\hat{R}|}, \quad \textit{Точность}^{+-} = \frac{|\hat{R} \cap R_{fuzzy2}|}{|\hat{R}|}$$

Проблема с точным совпадением:

$d_2 = \text{australie}$
$R_2 = \{ \text{ pays etranger, monde politique } \}$
$\hat{R}_2 = \{ \text{ benin, guinee-bissau, madagascar, moldavie, voirie de l'etat } \}$

"australie" $\leftrightarrow$ "monde politique" $\leftrightarrow$ "benin".

Результаты экспериментов:



$\textit{Точность}^{++} = 7\%, \quad \textit{Точность}^{+} = 35\%, \quad \textit{Точность}^{+-} = 46\%$

СПАСИБО ЗА ВНИМАНИЕ

Источники

Цитируемые источники:

1. B. Fortuna, M. Grobelnik and D.Mladenić, Visualization of Text Document Corpus , Informatica 29 (2005) 497–502
2. E. Ukkonen. (1995). On-line construction of suffix trees. *Algorithmica* **14**(3):249-260
3. D. Gusfield, Algorithms on strings, trees, and sequences : computer science and computational biology, 2007.
4. Manning C., Schütze H.(1999). «Foundations of Statistical Natural Language Processing». MIT Press. Cambridge, MA.

Используемые источники:

1. Филиппович Ю., Прохоров А. (2002) «Семантика информационных технологий: опыты словарно-тезаурусного описания».
2. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. (2003). «Формирование базы терминологических словосочетаний по текстам предметной области».
3. Б.В. Добров, Н.В. Лукашевич (2001). «Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе РОССИЯ».
4. Лукашевич Н.В. «Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России».

Источники

5. ГОСТ 7.25 - 2001. Тезаурус информационно-поисковый одноязычный
6. ISO 2788: Documentation - Guidelines for the establishment and development of monolingual thesauri. Second edition.
7. ANSI/NISO Z39.19-2005: Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies.
8. Gregory G. (1994) «Explorations in Automatic Thesaurus Discovery».
9. Cimiano P.(2006). «Ontology Learning and Population from Text». Algorithms, Evaluation and Applications.

11. Curran J., Moens M. (2002). «Improvements in automatic thesaurus extraction». Proceedings of the Workshop on Unsupervised Lexical Acquisition
12. Schutze H. (1998). «Automatic word sense discrimination». Computational Linguistics
13. Van der Plas L., Bouma G (2005). «Syntactic contexts for finding semantically related words». Proceedings of Computational Linguistics in the Netherlands 15.
14. Peirsman Y., Heylen K., Speelman D. (2006). «Putting things in order. First and second order context models for the calculation of semantic similarity»
15. Gregory G.(1993). «Automatic thesaurus generation from raw text using knowledge-poor techniques»