ТЕРМИНОЛОГИЧЕСКИЙ АНАЛИЗ ТЕКСТА НА ОСНОВЕ ЛЕКСИКО-СИНТАКСИЧЕСКИХ ШАБЛОНОВ

Ефремова Н.Э., Большакова Е.И.,
Носков А.А., Антонов В.Ю.

МГУ имени М.В. Ломоносова, факультет ВМиК

СОДЕРЖАНИЕ ДОКЛАДА

- Постановка задачи
- Термины и особенности их употребления
- Формализация особенностей
- Процедуры выявления терминов и их употреблений
- Стратегия совместного применения процедур

Используемые сокращения:
АОТ – автоматическая обработка текста
ПО – предметная область
ЕЯ – естественный язык
НТТ – научно-технический текст

ПОСТАНОВКА ЗАДАЧИ

- Решение многих задач АОТ требует выявления в текстах терминов

  Термин – слово или словосочетание, называющее понятие определенной ПО

  десятичная запятая, *донорно-акцепторная связь*

- Решение некоторых задач АОТ:
  - машинный перевод
  - литературно-научное редактирование

  требует выявление в отдельном тексте всевозможных **употреблений терминов**

  рентгеновское излучение – рентгеновские лучи, *излучение*

ПОДХОД К ВЫЯВЛЕНИЮ

- Обычно выявление терминов опирается на:
  - статистические особенности
  - лингвистические особенности
  - частичный синтаксический анализ
- Мы предлагаем учитывать:
  - типичную структуру терминов
  - варьирование отдельного термина
  - соединение нескольких терминов
  - характерные конструкции употребления терминов в НТТ
  - ✓ терминологический словарь ПО

ОСОБЕННОСТИ ТЕРМИНОВ

- типичная синтаксическая структура

  прил. + сущ. *электрический контур*
  сущ. + сущ. в род. падеже *тип данных*
  прил. + прил. + сущ. *слабая внешняя ссылка*

- терминологический словарь ПО:
  - **словарные** термины

  управление памятью, *первый закон Ньютона*
  - **новые** (**авторские**) термины

  тонкий клиент, *вимп*, *кэш второго уровня*

УПОТРЕБЛЕНИЕ В ТЕКСТЕ

- варьирование отдельного термина:
  одно понятие – **несколько** способов выражения

  алгебра логики – булева алгебра
  широкий атмосферный ливень – ШАЛ
- соединение нескольких терминов

  базовый класс ⊕ производный класс ⇒
  базовый и производный класс
- характерные конструкции:
  - определения авторских терминов

    Под конвейерным режимом понимают…
  - введения синонимов

    разрядностью, или длиной слова

ЛЕКСИКО-СИНТАКСИЧЕСКИЕ ШАБЛОНЫ

Для формализации выбран язык LSPL и его библиотека:

- язык позволяет описывать конструкции ЕЯ в виде **лексико-синтаксических шаблонов**
- библиотека реализует поиск по шаблонам описанных конструкций в тексте

Шаблоны фиксируют лексический состав и синтаксические связи формализуемых конструкций. Для этого используются:

- простые элементы **"базисом"**, **N**<базис,n=sing>**,** A N **<A=N>**
- сложные элементы **{**A**}** N**,** N1 **[**N2<c=gen>**],** A**|**Pa <Syn(N1,N2)>
- словарные условия
- имена шаблонов и параметры

  Term = {A} N1 [N2<c=gen>] <A=N1> **(N1)** ⇒ Term<c=ins>
- выделяемая конструкция

  Term1 "("Term2")" <Term1.c=Term2.c> # **Term1**

ПРИМЕРЫ ШАБЛОНОВ (1)

- Синтаксические образцы терминов:

  N1 A2 N2<c=gen> <A2=N2>

  технология двойной накачки

- Словарные термины:

  A1<битовый> {N1<массив> | N1<образ>}<1,1>

  битовый массив, *битовый образ*

- Контексты определения авторских терминов:

  Defin<c=acc> "будем" "называть" Term<c=ins> # Term

  Такие операции будем называть понятийными операциями

  "под" Term<c=ins> "понимается" Defin<c=nom> #Term

  Под продукцией понимается выражение…

ПРИМЕРЫ ШАБЛОНОВ (2)

- Правила образования лексико-синтаксических вариантов:

  N1 N2<c=gen> # *ввод данных*
  N1, *ввод*
  N1 N4<c=gen> <Syn(N2,N4)> *ввод информации*

- Соединения терминов:

  N1 N2<c=gen> "," N3<c=gen> {"и"|"или"} N4<c=gen> #
  N1 N2<c=gen>, N1 N3<c=gen>, N1 N4<c=gen>

  шинам адреса, данных и управления –
  шина адреса, шина данных, шина управления

- Контексты введения синонимов:

  Term1 "("Term2")" <Term1.c=Term2.c> # Term1, Term2

  взаимодействующих компонентов (подсистем)

ВЫЯВЛЕНИЕ ТЕРМИНОВ

- Набор процедур: каждая процедура – свой набор шаблонов
  - Термины-кандидаты – слова/словосочетания с типичной синтаксической структурой
- Вход: анализируемый текст, шаблоны
- Выявление терминов и их употреблений: поиск текстовых фрагментов, описываемых шаблонами
- Подсчет частоты
- Выход: термины с частотой употребления

ТЕСТИРОВАНИЕ ПРОЦЕДУР

- Процедуры по отдельности протестированы на НТТ из областей физики и информатики (объем ≈700 Кб)
- Использовались словари по физике (>3 тыс. терминов) и по информатике (>4 тыс. терминов)
- Оценивались полнота и точность выявления (в сравнении с экспертными списками):
  - **терминов**
  - их **употреблений** (вхождений в текст)

  Для синонимов и соединений: только полнота и точность выделения терминов, встретившихся в них

РЕЗУЛЬТАТЫ ТЕСТИРОВАНИЯ

Процедура	Выделение терминов		Выделение терминоупотреблений	
	полнота	точность	полнота	точность
Термины-кандидаты	58%	24%	54%	25%
Словарные термины	85%	94%	87%	95%
Авторские термины	67%	89%	70%	97%
Синонимы	57%	22%	–	–
Соединения	71%	30%	–	–

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ (1)

- Выявление терминов-кандидатов и соединений:
  - потеря точности

    крупный размер, *аналогичный результат*
  - потеря полноты

    индекс iCOMP, *обратная связь по релевантности*
- Выявление словарных терминов:
  - распознаны как термины общеупотребительные словосочетания или их части

    ряд – в ряде случаев, *за рядом исключений*

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ (2)

- Выявление авторских терминов и синонимов:
  - потеря полноты

    Регистр представляет собой совокупность...
  - словарные термины в контекстах определения

    Под прерыванием понимается сигнал...
- Выявление употреблений:
  - потеря полноты

    дисковый файл – файл на диске

    структурное и модульное программирование

ИДЕЯ ОБЪЕДИНЕНИЯ

- Расширение набора шаблонов:
  - повышается полнота, падает точность
  - требуется ручная работа
- Простое объединение списков терминов, выявленных процедурами:
  - повышается полнота, падает точность
- Учет процедурами списков терминов, выявленных другими процедурами:
  - повышается точность определения терминоупотреблений
  - выявленные из соединений термины давали прирост полноты выявления терминов на 12%

СТРАТЕГИЯ СОВМЕСТНОГО ПРИМЕНЕНИЯ ПРОЦЕДУР

(1) К тексту применяются процедуры выявления

(2) Словарные и авторские термины заносятся в S

(3) Термин-кандидат добавляется в S, если
его частью является словарный или авторский термин

(4) Пара синонимов добавляется в S,
если один из них уже в S

(5) Термины из соединений добавляются в S, если
среди них есть разрывный термин из S (или словарный)

(6) Для терминов из S ищутся лексико-синтаксические варианты и добавляются в S

(7) В S добавляются термины-кандидаты
с частотой выше некоего порога

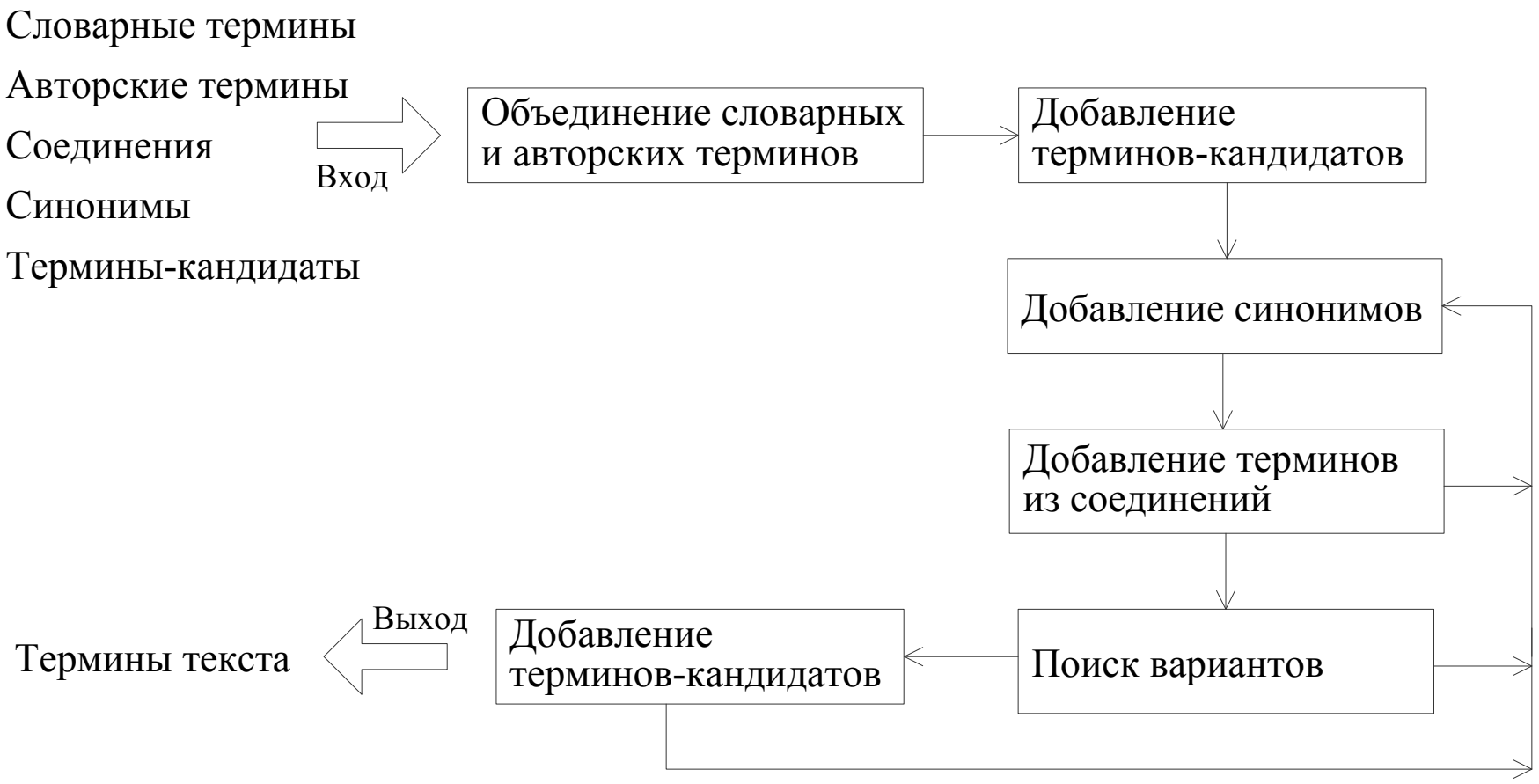
(8) Повторяем шаги, начиная с 3

РЕЗУЛЬТАТЫ ПРИМЕНЕНИЯ СТРАТЕГИИ

- Для оценки результатов использовалась F-мера:
  F = 2 x Precision x Recall / (Precision + Recall)
- Сравнивались списки терминов, полученные:
  - простым объединением списков терминов, выявленных процедурами
  - применением стратегии
- В среднем прирост:
  - F-меры выявления терминов – 10%
  - F-меры выявления терминоупотреблений – 7%
- Проблемы:
  - как термины выявляются общенаучные словосочетания (*различные цели*)
  - один вариант связывается с несколькими терминами (*регистр адреса*, *регистр команды* – *регистр*)

СПАСИБО ЗА ВНИМАНИЕ!


Словарные термины
Авторские термины
Соединения
Синонимы
Термины-кандидаты
Вход
словарные и авторские термины
S1
Объединение словарных и авторских терминов
S1
Термины-кандидаты
S2
Добавление терминов-кандидатов
S3
S2
Синонимы
Добавление синонимов
S4
S3
Термины из соединений
Добавление терминов из соединений
Если S4□S3 (S2:=S4)
Если S5□S4 (S2:=S5)
Термины текста
Выход
Добавление терминов-кандидатов
Поиск вариантов
S6
S5
Термины-кандидаты
S5
S4
Варианты
Если S6□S5 (S2:=S6)


Словарные термины
Авторские термины
Соединения
Синонимы
Термины-кандидаты
Вход
Объединение словарных и авторских терминов
Добавление терминов-кандидатов
Добавление синонимов
Добавление терминов из соединений
Поиск вариантов
Добавление терминов-кандидатов
Выход
Термины текста