Сравнительный анализ статистических алгоритмов синтаксического анализа на основе деревьев зависимостей

Казенников А.О.
ИППИ им. А. А. Харкевича РАН
Лаборатория №15

Задача

Оценить алгоритмы синтаксического анализа по отношению к системе ЭТАП-3:

- Алгоритмы на основе максимальных остовных деревьях(MST)
- Алгоритмы на основе систем переходов(TS)

Сложности прямого сравнения

- Разная постановка задач
- Разные входные данные
  - Пунктуация
  - Морфологические признаки
  - Омонимия слов предложения
- Разные выходные данные
  - Именованные связи

Корпус SynTagRus

- ~500 тыс. слов
- ~35 тыс. предложений
- ~15.4 слова на предложение
- Слово:
  - Словоформа
  - Лемма
  - Морфологические характеристики
  - Хозяин
  - Имя связи

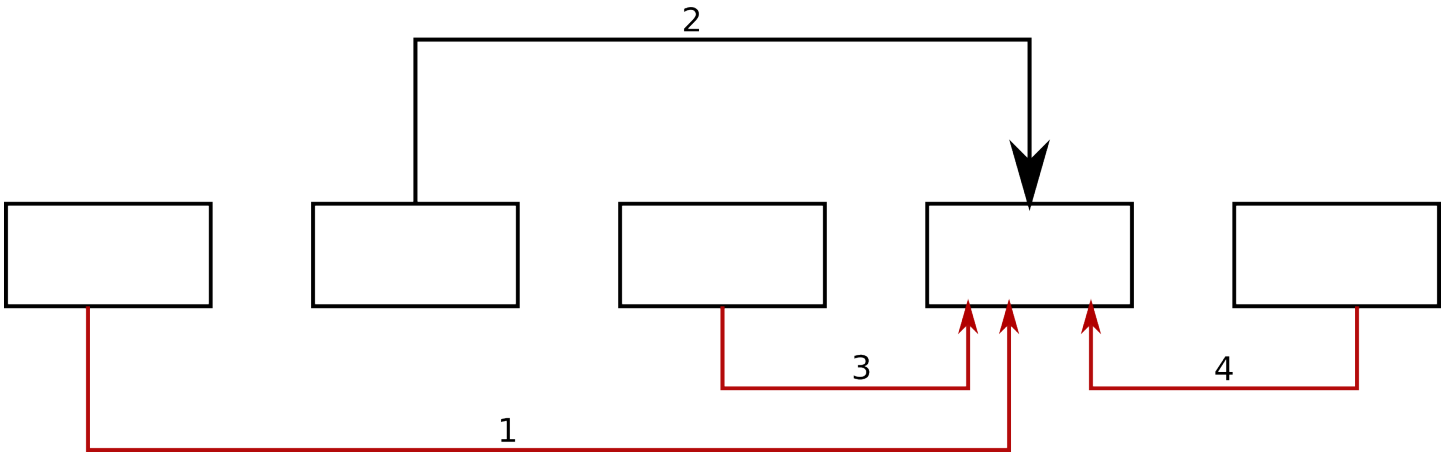
MST. Общая идея

- На входе: полный граф связей предложения
- Каждой связи приписывается вес с помощью функции оценки связи
- Построение дерева из графа с максимальной суммой оценок связей
- Два уровня:
  - Алгоритм построения функции оценки веса связи
  - Алгоритм построения дерева на основе весов связей

MST. Оценка связи

- Производится независимо от других связей
- Использует только «линейный» контекст
- S(эталон) > S(неэталон)
- Ранжирование потенциальных связей
- Алгоритм - SVMRank

MST. Получение данных для обучения



- Для каждого слова — ~*n* примеров
- 2 > 1, 2 > 3, 2 > 4
- 1 ? 3 ? 4

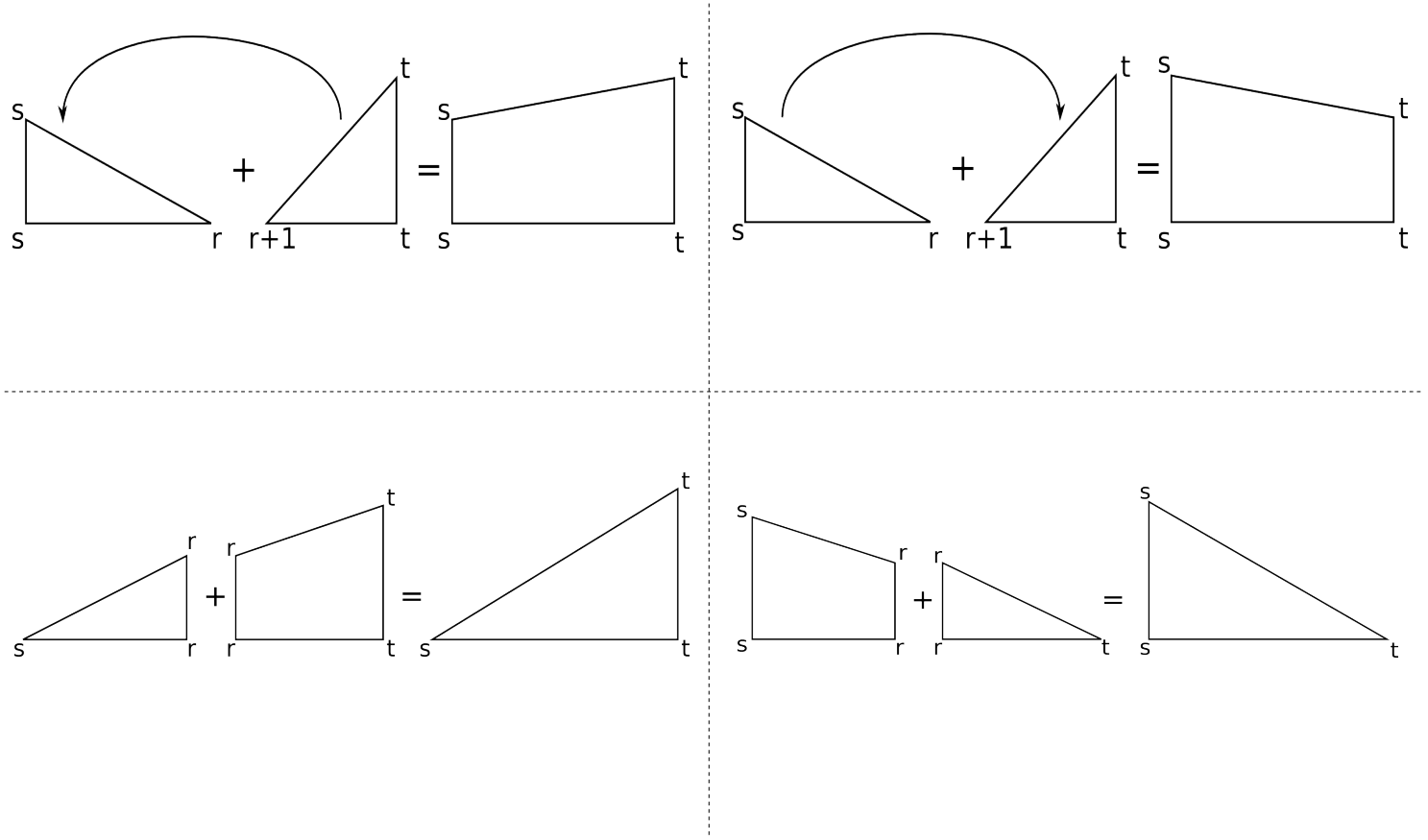
MST. Построение дерева

- Два алгоритма:
  - Проективный. Табличный алгоритм вида CKY
  - Непроективный. Графовый алгоритм на основе CLE
- Оценка каждой связи проводится независимо от других

Алгоритм Эйснера

- Табличный алгоритм CKY-типа

  Структура ячейки таблицы:



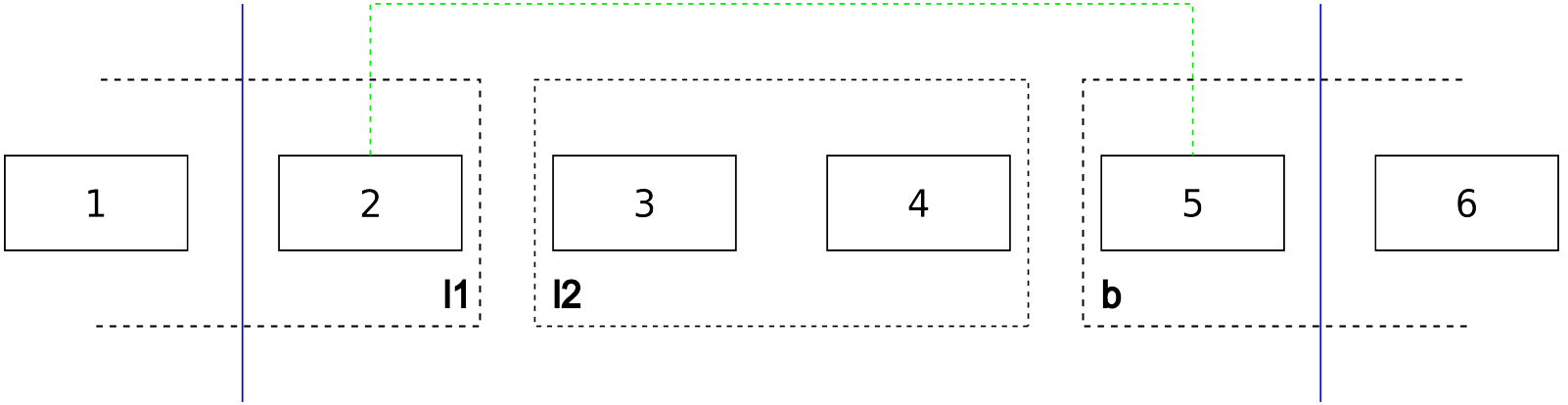
MST. Проблемы подхода

- Квадратичная зависимость данных от длины предложения
- Сильно несбалансированные данные для обучения
- Не учитывается «древесный» контекст
- 3 уровня оценки:
  - Связь - 2.5% ошибок
  - Хозяин слова - ~20% ошибок
  - Предложение - ~80% ошибок

Синтаксический анализ на основе системы переходов

- Один проход по предложению слева направо
- Использование истории принятых решений
- 3 составляющих:
  - Конфигурация — состояние системы разбора
  - Набор возможных действий
  - Алгоритм преобразования эталонной структуры в набор действий
- Алгоритм — разбиение по нескольким классам

TS. Элементы



- Конфигурация
- Набор действий
- Эталон → действия

- LeftArc
- RightArc
- NoArc
- Shift

TS. Получение данных для обучения

- Отсортировать слова по возрастанию номера правого слова связи
- Отсортировать по убыванию расстояния до правого слова связи
- Проводить связь
- Или использовать NoArc/Shift для подбора левого/правого слова

TS. Особенности подхода

Преимущества:

- Использование частично построенное дерева для выбора следующего действия
- Простой алгоритм обучения

Недостатки:

- Низкая точность проведения длинных связей
- Накапливание ошибок с ростом длины предложения

Данные для экспериментов

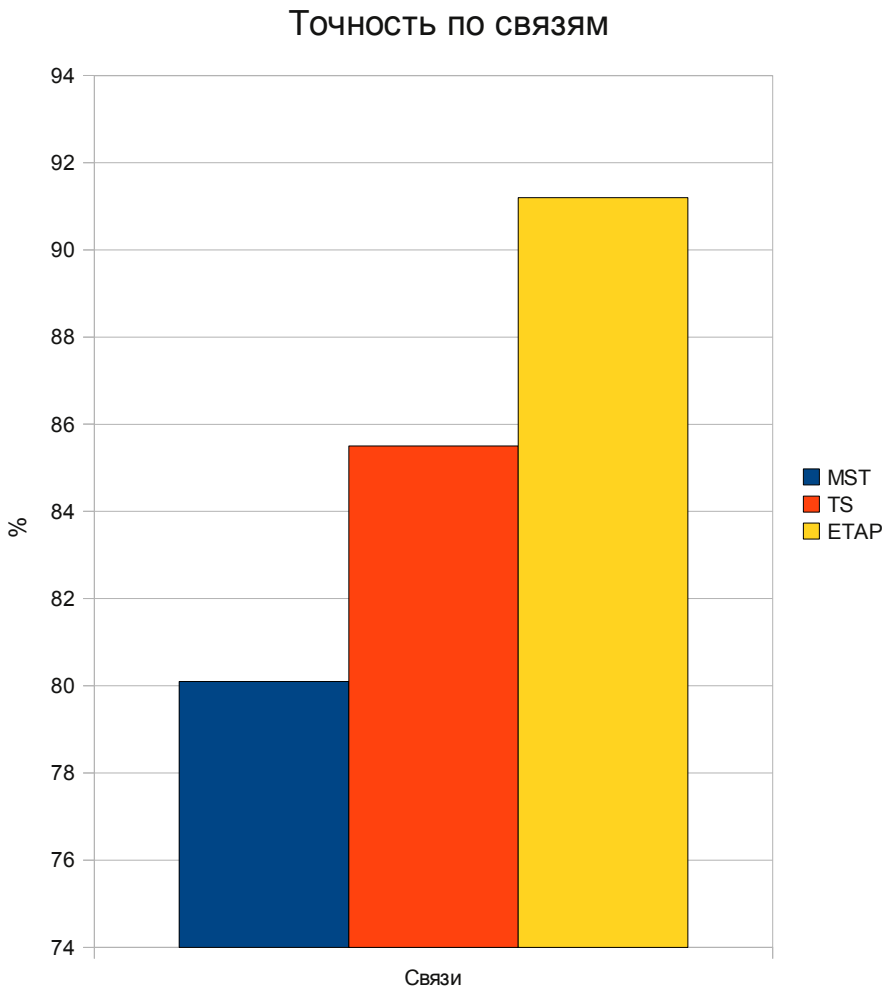
Выборка для обучения:

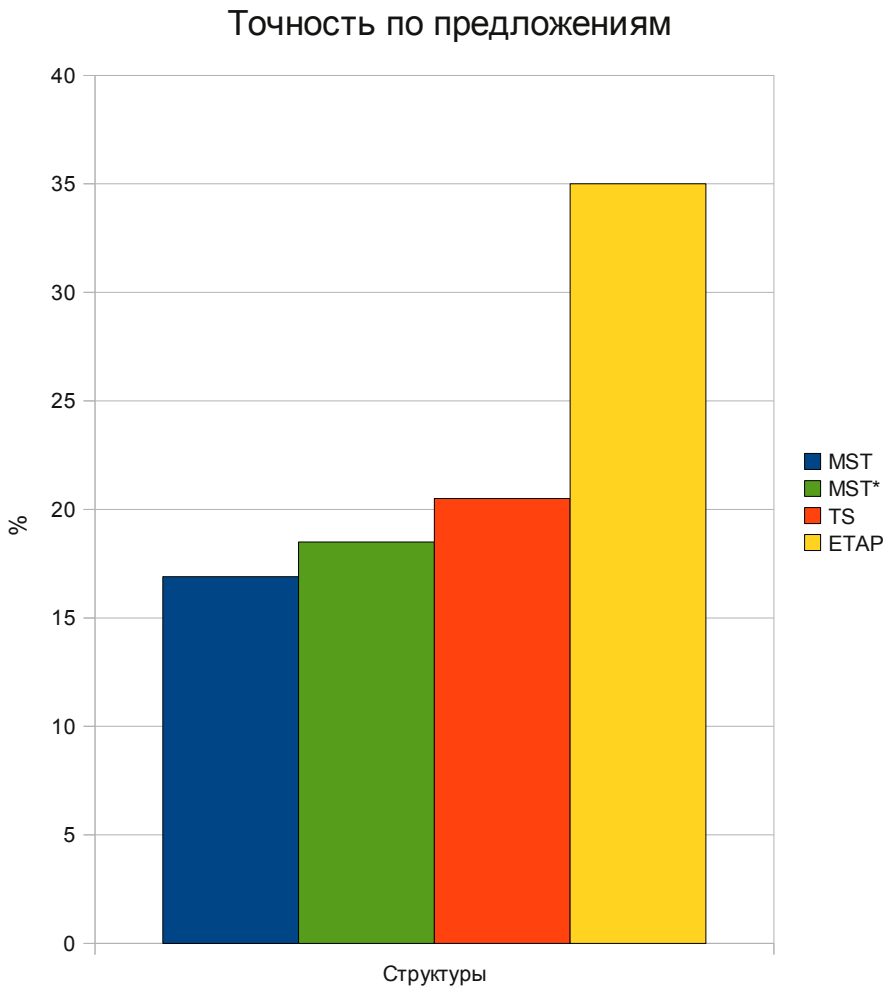
- MST — 5 тыс. предложений
- TS — 15 тыс. предложений

Проверочная выборка:

- 5 тыс. предложений

Результаты


Точность по связям
94
92
90
88
86
84
82
80
78
76
74
%
Связи
MST
TS
ETAP


Точность по предложениям
40
35
30
25
20
15
10
5
0
%
Структуры
MST
MST*
TS
ETAP

Спасибо за внимание!