Постановка задачи

Решаемая задача: сегментация библиографических ссылок

- алгоритм не должен быть привязан к какому-то одному формату цитирования
- пользователь должен иметь возможность изменять правила сегментации (не только формат полей, но и их состав)

Предлагаемое решение

Программа на Java

- на входе:
  - файл со ссылками
  - пользовательский файл конфигурации (правила сегментации)
- на выходе: файл формата .json с объектами, соответствующими ссылкам и содержащими пары название поля – значение поля

► *См. примеры на с. 5.*

Файл конфигурации: синтаксис

Содержит элементы КС-грамматики.

Структура:

1. определения элементов (полей и их сегментов, разделителей, вспомогательных символов)
   - в левой части – наименование элемента
   - в правой части – правило деривации из ранее определенных элементов (нетерминалов) и регулярных выражений (терминалов)
     Ограничения:
     - только операция конкатенации
     - нерекурсивная грамматика: нельзя определять элемент через самого себя
2. список полей (`FIELDS`)
3. определения ссылок (`ENTRIES`)

► *См. пример на с. 6.*

Реализация алгоритма

1. поэтапное (снизу вверх) построение регулярных выражений, соответствующих полям
2. построение массива регулярных выражений, описывающих типы ссылок, с сопровождающей информацией о содержащихся полях
3. сравнение ссылок с регулярными выражениями, выбор оптимального варианта для каждой ссылки, извлечение полей и запись объекта в выходной файл

Кирилл Михайлович Карпенко
Автоматическая обработка и преобразование библиографических ссылок
Курсовая работа (бакалавриат, III курс)
Научный руководитель: Павел Валерьевич Гращенков

Оценка качества

- Создан файл конфигурации для затекстовых ссылок в формате ГОСТ 7.0.5—2008 (часть типов)
- Создан тестовый корпус (91 ссылка, 535 элементов)

Таб. 3. Значения метрик для тестового корпуса

	n	Распознано всего	Распознано корректно	p (точность)	r (полнота)	F1-мера
Элементы	535	483	463	0,96	0,87	0,91
Полные ссылки	91	84	64	0,76	0,70	0,73

Анализ результатов

Практически все ошибки связаны с неполнотой или дефектами файла конфигурации:

- все 7 нераспознанных ссылок + 8 из 17 некорректно отсегментированных ссылок: типы, не определенные в файле
  - электронные ресурсы:
    Национальный корпус русского языка // http://www.ruscorpora.ru.
  - доклады на конференциях:
    Толдова С. Ю. Дифференцированное маркирование прямого дополнения в мокшанском языке. Доклад на Международном совещании по дифференцированному маркированию актантов, Институт языкознания РАН, 22—23 апреля 2016.
  - неопубликованные работы:
    Афанасьева Т. И. Русские переводы конца XIV века // Сборник статей памяти В. М. Живова. М. (в печати).
  - ссылки со сведениями об ответственности:
    Новый объяснительный словарь синонимов русского языка / Под общ. рук. акад. Апресяна Ю. Д. 2-е изд. М.: Языки славянской культуры, 2003.

- неточное определение формата внутри типа ссылки
  - неучтённая опциональность поля
    нет издательства:
    Сердобольская Н. В. К типологии выражения значения генерического события в конструкциях с сентенциальными актантами // Доклад на восьмой конференции по типологии и грамматике для молодых исследователей. Санкт-Петербург, 2011.
  - дефекты в определениях полей (неучтённые варианты):
    фамилии из нескольких слов с разделителем – пробелом: *Ле Гийю де Пенанрос Э.*
    полные личные имена: *Зализняк Анна А.*
    другие неучтенные подстроки: *Vol.* – том

2 случая – опечатки (пропуск разделителя).
Ошибки сегментации при наличии опечаток на границах полей закономерны: успех сегментации в этой программе часто зависит от пограничных участков соответствующих регулярных выражений.

Перспективы

- Имплементация рекурсивной грамматики: позволит создавать намного более сложные правила
- Совмещение с вероятностным подходом: позволит частично обрабатывать некорректно форматированные ссылки

Кирилл Михайлович Карпенко
Автоматическая обработка и преобразование библиографических ссылок
Курсовая работа (бакалавриат, III курс)
Научный руководитель: Павел Валерьевич Гращенков

Рис. 1. Запуск программы в командной строке Windows 8.1 (пример)

Таб. 1. Фрагменты входного и выходного файлов (примеры)

Примеры ссылок во входном файле	Пример объектов в выходном .json-файле
Латышева А. Н. О семантике условных, причинных и уступительных союзов в русском языке // Вестн. Моск. ун-та. Сер. 9. Филология. 1982. No 5. С. 51—59.	{ "volume": "", "pagination": "", "issue": "No 5", "year": "1982", "collection-editors": "", "collection-title": "Вестн. Моск. ун-та. Сер. 9. Филология", "article-title": "О семантике условных, причинных и уступительных союзов в русском языке", "authors": "Латышева А. Н." }
Подлесская В. И. Условные конструкции: стратегии кодирования и функциональная мотивация // Тестелец Я. Г., Рахилина Е. В. (ред.) Типология и теория языка: от описания к объяснению (К 60-летию А.Е.Кибрика). Москва: Языки русской культуры, 1999. С. 255—273.	{ "pagination": "", "city": "Москва", "year": "1999", "collection-editors": "Тестелец Я. Г., Рахилина Е. В. (ред.)", "collection-title": "Типология и теория языка: от описания к объяснению (К 60-летию А.Е.Кибрика)", "publishers": "Языки русской культуры", "article-title": "Условные конструкции: стратегии кодирования и функциональная мотивация", "authors": "Подлесская В. И." }
Кузнецова А. И., Хелимский Е. А., Грушкина Е. В. Очерки по селькупскому языку. Тазовский диалект. Том I. М.: Изд-во Московского университета, 1980.	{ "book-title": "Очерки по селькупскому языку. Тазовский диалект", "volume": "Том I", "pagination": "", "city": "М.", "year": "1980", "publishers": "Изд-во Московского университета", "authors": "Кузнецова А. И., Хелимский Е. А., Грушкина Е. В." }

Кирилл Михайлович Карпенко
Автоматическая обработка и преобразование библиографических ссылок
Курсовая работа (бакалавриат, III курс)
Научный руководитель: Павел Валерьевич Гращенков

Таб. 2. Фрагмент файла конфигурации (пример)

```
authors = "[\p{Lu}][\p{Ll}]*([-—\-][\p{L}]+)* [\p{Lu}][\p{Ll}]?\.([ —\-
]?[\p{Lu}][\p{Ll}]?\.)*(, [\p{Lu}][\p{Ll}]*([-—\-][\p{L}]+)*
[\p{Lu}][\p{Ll}]?\.([ —\-]?[\p{Lu}][\p{Ll}]?\.)*)*( \((отв\.
?)?ред\.\))?"
city = "([\p{Lu}][\p{L} —\-]*\.?)(, [\p{Lu}][\p{L} —\-]*\.?)*"
year = "\d{4}"
volume = "(Т(ом)?\.? (([IVX]+)|(\d+))?)?"
issue = "(((Вып\.)|(No\.?)|№) ?\d+)?"
semicolon = "\:"
comma = ","
...
FIELDS
authors
book-title
city
publishers
year
article-title
collection-title
collection-editors
...
issue
ENTRIES
article-in-journal = authors period space article-title period space slash
space collection-editors period space collection-title period space year
period space volume period space issue period space pagination period
article-in-book = authors period space article-title period space slash
space collection-editors period space collection-title period space city
semicolon space publishers comma space year period space pagination period
book = authors period space book-title period space volume period space
city semicolon space publishers comma space year period space pagination
period
...
```