

Аннотирование по запросу: ‘связность или информативность’?

Кондратьев Михаил

Контекстное аннотирование

- **Идеальная аннотация по запросу:**

фрагмент связного текста, описывающего информацию исходного документа, соответствующую запросу пользователя.

- **Создание аннотации из фрагментов документа:**

- большие фрагменты имеют хорошую связность но мало информативны

- фрагменты малого размера содержат много информации но уменьшают связность

Как строить аннотации

Открытые вопросы:

- каков должен быть размер фрагментов, составляющих аннотацию?
- сколько фрагментов должны составлять аннотацию?
- каким должен быть вид фрагментов?

Цель работы: экспериментально оценить влияние информативности и связности на качество аннотирования

Подход 1: единый фрагмент

- **Мотивация:**

единый фрагмент текста обладает наилучшей связностью и хорошо описывает требуемую информацию

- **Построение аннотации:**

связный фрагмент, содержащий максимальное количество термов запроса

- **Недостатки:**

возможная низкая информативность

Подход 2: комбинация предложений

- **Мотивация:**

комбинация предложений содержит больше термов запроса и по крайней мере локально связна

- **Построение аннотации:**

упорядоченное множество предложений, количество уникальных термов в котором максимально

- **Недостатки:**

потеря в связности по сравнению с единым фрагментом

Подход 3: фрагменты предложений

- **Мотивация:**

использование максимально возможного количество термов запроса позволяет достичь макс. информативности

- **Построение аннотации:**

упорядоченное множество фрагментов предложений, количество уникальных термов в котором максимально

- **Недостатки:**

самая низкая связность из 3 прогонов

Анализ результатов

1. Эксперименты не выявили явного лидера среди выполненных прогонов
2. Прогон 3 показал максимальные среди всех участников значения PrecisionAnnotation
 - 10 из 12 оценок для прогона 3 среди участников
 - прогоны 1 и 2 показали высокие результаты
3. Прогон 1 показал наихудшие значения по метрике AnnotationError (12 из 12 оценок)
4. Прогон 3 показал наилучшие значения по метрике AnnotationError (11 из 12 оценок)
5. Прогоны равное число раз показывали лучший результат по метрике AnnotationAccuracy

Выявленные недостатки

1. Неоптимальное использование доступного объема аннотации

запрос: 'футбольные турниры в Москве'

аннотация: 'В первый день турнира наставник "Старта" Юрий Фокин... ' (51 символ)

2. Неверное определение границ предложения

пример аннотации: 'считается днем принятия Конституции Российской Федерации. Одновременно прекращается действие...'