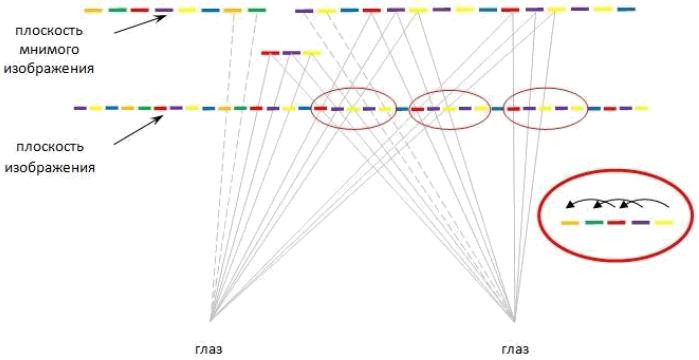


Рис. 3. Пример сдвига на две позиции

Выводы

Рассмотренные методы позволяют совершенствовать технологию формирования растровых стереоизображений, а разработанный плагин дает возможность ускорения этого процесса в графическом редакторе GIMP.

Литература.

1. ProGIMP – сайт о Гимп. GIMP – The GNU Image Manipulation Program / Интернет-ресурс. – Режим доступа: www / URL: http://www.progimp.ru/news/site/ [текст].

2. Scheme / Интернет-ресурс. – Режим доступа: www / URL: http://ru.wikipedia.org/wiki/Scheme [текст].

3. Стереоизображение – это просто / Интернет-ресурс. – Режим доступа: www / URL: http://habrahabr.ru/post/127681/ [изображения].

Сарры Н.А.

Науч. рук. к.т.н., доц. Звенигородский А.С.

Донецкий национальный технический университет

Определение биграмм на материале научных текстов по извлечению данных из текстов

В данном докладе рассматривается извлечение информации о предметной области научных текстов, что является неотъемлемой частью задачи выделения важных терминов. В качестве предметной области была выбрана область, связанная с извлечением данных из текстов, большинство терминов которой являются не однословными. Не однословные термины характеризуются термином коллокация.

Коллокация – неслучайное сочетание двух и более лексических единиц, характерное как для языка в целом, так и для определенного типа текстов. Использование статистических мер позволяет выделять из текста коллокации и ранжировать их по степени устойчивости в соответствии со значениями выбираемых мер [1].

Для текстов научного стиля статистически определяются составные слова и устойчивые конструкции, характеризующие особенности стиля, смысловую и коммуникативную структуру текста.

В основу статьи следующие гипотезы [2]:

1. Использование меры *MI* позволяет выделить ключевые не однословные термины, которые характеризуют предметную область.

2. Использование меры *t-score* позволяет выделить устойчивые сочетания, устойчивые конструкции, характеризующиеся стилистическими особенностями научного текста.

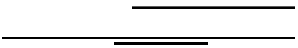
Статистические мера MI – Mutual Information (коэффициент взаимной информации) [2] определяется по формуле (1):

_____ , 1)

где: *n* – ключевое слово; *c* – коллокат; *f(n,c)* – абсолютная частота встречаемости ключевого слова *n* в паре с

коллокатом *c*; *f(n), f(c)* – абсолютные частоты ключевого слова *n* и слова *c* в корпусе; *N* – объем корпуса (количество словоупотреблений) [2].

Мера t-score [2] определяется по формуле (2):



2)

С точки зрения теории вероятности, мера MI является способом проверить независимость появления двух слов в тексте – если слова полностью независимы, то вероятность их совместного появления равна произведению вероятностей появления каждого из них.

Мера t-score используется гораздо реже, чем мера MI, поскольку она является лишь несколько модифицированным ранжированием коллокаций по частоте. Очевидно, что значение данной меры тем выше, чем выше частота коллокации в наборе текстов. Данная мера содержит коррекционный компонент, но эта поправка отражается лишь на самых частотных словах.

Был подобран набор текстов в области извлечения данных. На основании обработки этих текстов была получена предварительная информация о терминах, употребляемых в текстах, посвященных извлечению данных.

В табл. 1 представлен список биграмм, полученных с помощью меры MI. Этого списка достаточно, чтобы получить предварительную информацию о наиболее важных не однословных терминах: объектах исследования, материале, методах, результатах.

Таблица 1 – Биграммы (MI-score), выделяющиеся как для лексем, так и для словоформ

№	биграмма	
1	лексическая	единица
2	математическая	лингвистика

3	семантический	анализатор
4	морфологическая	разметка
5	научная	статья
6	предметная	область
7	анализ	текста
8	выделение	сущностей
9	автоматическое	извлечение
10	извлечение	информации
11	профессиональный	словарь
12	целевой	фрейм
13	фильтрация	документа
14	обучающая	выборка
15	шаблоны	фраз

Используя меру *t-score* можно выделить те сочетания, которые могут рассматриваться как терминологические. Таким образом, был получен список биграмм общий для всех текстов из набора (см. табл. 2).

Данное исследование показывает, что:

– использование меры MI позволяет выделить «ключевые» не однословные термины, характеризующие предметную область набора текстов;

Таблица 2 – Терминологические биграммы (t-score), выделяющиеся как для лексем, так и для словоформ

№	Лексемные биграммы	
1	лексическая	единица
2	математическая	лингвистика
3	семантический	анализатор
4	выделение	сущностей
5	извлечение	информации
6	фильтрация	документа
7	модель	текста

– использование меры t-score позволяет выделить: «устойчивые сочетания», «устойчивые конструкции», характеризующие стилистические особенности научных текстов, коллокации, общие для всех текстов из набора.

Результаты исследования являются основой для разработки алгоритмов определения принадлежности текстов к научной тематике по извлечению данных.

Литература.

1. Ягунова Е.В. Формальные и неформальные критерии вычленения ключевых слов из научных и новостных текстов / Е.В. Ягунова. – М. – 2010. – С. 340 – 355.

2. Ягунова Е.В., Пивоварова Л.М. Извлечение и классификация коллокаций на материале научных текстов. Предварительные наблюдения / Е.В. Ягунова, Л.М. Пивоварова. – СПб. – 2010. – С. 356-364.