Обзор исследований в рамках Cross-Language Evaluation Forum в 2006 году

© М.С. Агеев

Научно-исследовательский вычислительный центр
МГУ им. М.В.Ломоносова,
АНО Центр информационных исследований
ageev@mail.cir.ru

Аннотация

В статье дается обзор задач и исследований, проводимых в рамках международного семинара Cross-Language Evaluation Forum в 2006 году.

1. Введение

Ежегодно, на очной встрече семинара РОМИП, происходит обсуждение планов по развитию РОМИП. При этом полезно учитывать опыт наших коллег — Cross-Language Evaluation Forum (CLEF).

CLEF является ежегодным международным семинаром, посвященным экспериментальному изучению методов информационного поиска, с упором на задачи многоязычного поиска и задачи интеграции многоязычных коллекций, в первую очередь — для европейских языков.

Под многоязычным поиском понимается комплекс задач информационного поиска, в которых вопрос пользователя к информационной системе формулируется на одном языке, а коллекция документов содержит документы на различных языках.

Целями CLEF [2] является продвижение исследований по созданию многоязычных информационных систем, проведение экспериментов по всем типам задач многоязычного информационного поиска, объединение усилий исследователей для разработки многоязычных систем следующего поколения.

Задачи создания многоязычных информационных систем особенно актуальны для европейского сообщества в связи с растущей интеграцией разных культур в рамках единого экономического и политического пространства Евросоюза.

В 2006 году состоялся седьмой семинар CLEF, собравший 90 участников (коллективов) со всех континентов мира. В рамках CLEF'2006 проведены 8 дорожек для различных задач информационного поиска. Задания сформулированы для более 20 различных языков мира, включая западноевропейские, азиатские (китайский, японский, хинди, и др.), русский, а также экзотические (оромо, телугу) языки.

Коллектив НИВЦ МГУ поддерживает партнерские отношения с CLEF. Мы способствовали предоставлению легальных русскоязычных коллекций документов для CLEF, проводили оценку релевантности документов для русского языка. Н.В. Лукашевич является членом наблюдательного совета конференции CLEF. В 2005 году мы приняли участие в CLEF [9]. В этом году автор принял участие в очном семинаре CLEF 2006 Workshop [1].

2. Организация CLEF

CLEF координируется [1] Итальянским институтом Науки и Информационных Технологий.

Как и в РОМИП, структура годового цикла CLEF состоит из следующих этапов:

1) согласование списка и правил дорожек, сбор коллекций документов;
2) регистрация участников, подписание соглашений и распространение коллекций;
3) выполнение заданий участниками, отправка результатов;
4) оценка результатов;
5) очная встреча.

Даты проведения этапов зависят от дорожки. Очная встреча проходит в рамках семинара CLEF Workshop, присоединенного к Европейской Конференции по Электронным Библиотекам (ECDL), в конце сентября (в этом году — 20-22 сентября в Аликанте, Испания).

Все результаты участников, а также обзоры по каждой дорожке, публикуются на сайте CLEF [1], графики приводятся для лучших 5 результатов. Обсуждается перспектива открытого опубликования детальных (по каждому топику) таблиц результатов участников.

Традиционно, большинство участников CLEF — из Европы, но также принимают участие исследователи с других континентов (рис. 1).

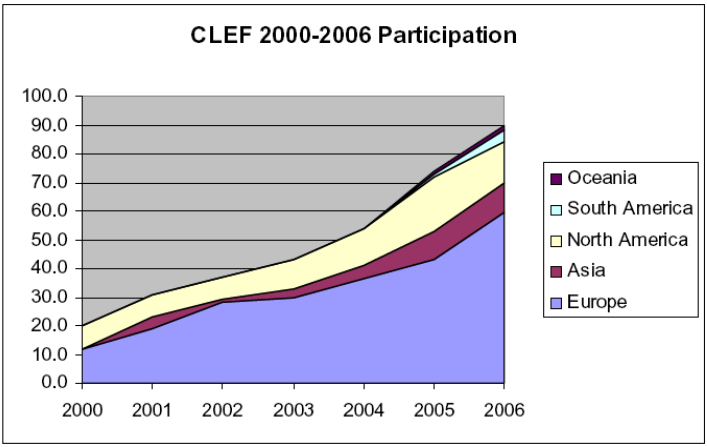

Рис. 1 Количество участников CLEF 2000-2006, с распределением по континентам [2]

3. Обзор дорожек CLEF 2006

Изложение этого раздела следует статье [2].

В CLEF 2006 участникам были предложены следующие дорожки:

1) моноязычный, двуязычный и многоязычный поиск документов (Ad Hoc);
2) моноязычный и многоязычный поиск по структурированным документам в области социальных наук (Domain-Specific);
3) интерактивный многоязычный поиск (iCLEF);
4) многоязычный поиск ответов на вопрос (QA@CLEF);
5) многоязычный поиск по коллекции изображений (ImageCLEF);
6) многоязычный поиск по коллекции аудиозаписей речи (CL-SR);
7) многоязычный поиск web-документов (WebCLEF)
8) многоязычный поиск с учетом географических отношений (GeoCLEF)

Задача *моноязычного поиска* заключается в поиске документов в коллекции на том же языке, на котором сформулирован запрос.

Для *двуязычного поиска* коллекция документов и вопросы сформулированы на разных языках (query language vs. target language).

Для *многоязычного поиска* предлагается коллекция запросов на нескольких различных языках и/или коллекция документов на нескольких различных языках.

На рис. 2 показано распределение количества участников для различных дорожек, с 2000 по 2006 годы.

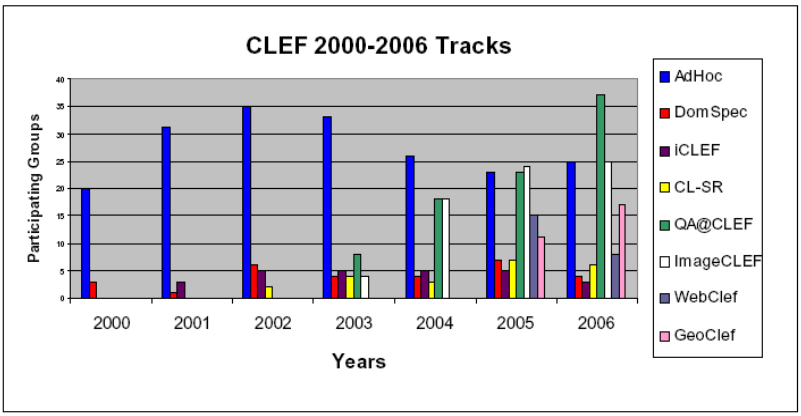

Рис. 2 Количество участников для различных дорожек CLEF, с 2000 по 2006 годы [2]

3.1. Многоязычный поиск документов (Ad Hoc)

Для многоязычного поиска в качестве коллекций документов использовались сообщения новостных агентств на французском, португальском, болгарском и венгерском языках. Задания были сформулированы на болгарском, английском, французском, немецком, венгерском итальянском, португальском и испанском языках.

Для двуязычного поиска были предложены задания поиска по англоязычной коллекции документов, с заданиями на языках: амхарском, китайском, хинди, индонезийском, оромо, телугу.

В 2006 году впервые в рамках CLEF была проведена дорожка **«robust task».** Эта дорожка отличается от стандартного Ad Hoc использованием другой усредненной метрики оценки качества поиска: в отличие от стандартной метрики «среднее значение Average Precision [10] по всем заданиям (MAP)» используется «среднее *геометрическое* значение Average Precision по всем заданиям (GMAP)».

$$\mathrm{MAP} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{AvgPrec}\left(\mathrm{topic}_i\right), \qquad \mathrm{GMAP} = \left(\prod_{i=1}^{N}\mathrm{AvgPrec}\left(\mathrm{topic}_i\right)\right)^{\frac{1}{n}}$$

Метрика GMAP подчеркивает важность получения стабильно хорошего результата по всем заданиям, в отличие от хорошего *в среднем* результата для стандартной метрики MAP.

3.2. Поиск по структурированным документам в области социальных наук (Domain-Specific)

В данной дорожке предлагались задания по моноязычному, двуязычному и многоязычному поиску для коллекций:

1) параллельных (переведенных) немецко-английских текстов по социальным наукам;
2) русскоязычных текстов по социальным наукам (библиографические описания ИНИОН и Соционет (http://www.socionet.ru)).

Также предлагались лингвистические ресурсы — двуязычные словари и тезаурусы по социальным наукам, которые можно использовать не только для перевода терминов, но и для улучшения алгоритмов оценки релевантности документов.

3.3. Интерактивный многоязычный поиск (iCLEF)

Участники дорожки iCLEF представили для оценки системы многоязычного поиска по коллекции изображений.

В качестве коллекции изображений использовалась база Flickr (www.flickr.com) — динамичная и быстро развивающаяся база данных изображений с текстовыми заголовками, аннотациями и комментариями на различных языках. Использовался срез базы с несколькими миллионами изображений.

Пользователям было предложено три задания, которые нужно было выполнить в интерактивном режиме за ограниченное время:

1) Найти как можно больше зданий парламентов европейских стран, картинки как изнутри так и снаружи здания.
2) Найти пять иллюстраций к тексту «история шафрана» (текст прилагался).
3) Найти название пляжа, на котором отдыхает этот краб. (Картинка с крабом, отдыхающем на песке, прилагалась. Название пляжа имелось в аннотации на немецком языке к одной из картинок в базе, на которой был изображен данный краб, но пользователю, конечно, заранее ничего не сообщали.)

3.4. Многоязычный поиск ответов на вопрос (QA@CLEF)

Основная задача заключалась в том, что система должна найти

1) Точный ответ на заданный запрос.
2) Сниппет из текста документа, который подтверждает данный ответ.

Участникам было предложено исполнить 200 запросов следующих типов [3]:

1) Поиск фактов, событий. Например: *Кто был президентом США в 1962 году? Какой партии принадлежал Гитлер?*
2) Поиск определений для людей, организаций, предметов. Например: *Что такое роутер? Кем является Лиза Мария Пресли?*
3) Найти список людей или объектов. Например: *Назовите книги Жюля Верна*.

Кроме основной задачи — поиска точных ответов на заданный вопрос, предлагались также задания, новые для CLEF:

Поиск в Википедии (wikipedia.org) информации, новой для заданной статьи. На вход системе подается статья из википедии. Нужно найти факты, которые релевантны теме данной статьи, но не отражены в ней. Такие факты могут содержаться

a) в другой (возможно, связанной по ссылкам) статье на том же языке;
b) в аналогичной или связанной статье на другом языке.

Задание по проверке ответов: на вход системе передается вопрос, ответ некоторой вопросно-ответной системы и сниппет, выданный в подтверждение данного ответа. От системы требуется ответить, действительно ли сниппет подтверждает ответ.

В качестве basic line для оценки алгоритма используется система, которая всегда дает ответ «Да». Стоит отметить, что не всем алгоритмам удалось улучшить этот basic line.

3.5. Многоязычный поиск по коллекции изображений (ImageCLEF)

Для этой задачи предлагалось несколько коллекций аннотированных изображений.

Рассматривались задачи:

- поиска изображений по запросу, состоящему из текста;
- поиска изображений по запросу, состоящему из текста плюс изображения;

- автоматического аннотирования изображений.

Часть заданий требовало обязательного использования алгоритмов обработки изображений, но были также задания, которые не требовали обработки изображений. Базовая система обработки изображений была предоставлена участникам.

3.6. Многоязычный поиск по коллекции аудиозаписей речи (CL-SR)

Дорожка многоязычного поиска по коллекции речевых аудиозаписей включала две задачи по различным коллекциям [4]:

1. Поиск релевантных сегментов в коллекции англоязычных интервью. Для данной коллекции предоставлены следующие метаданные:
   - проставленные вручную границы времени тематически однородных фрагментов интервью;
   - результаты системы автоматического распознавания речи, обработанные одной из лучших известных программ распознавания речи (уровень ошибок около 25%).

В качестве заданий для поиска были предоставлены топики на нескольких европейских языках.

2. Поиск релевантных сегментов в коллекции интервью на чешском языке. Для данной задачи нужно было автоматически найти время начала релевантных сегментов интервью.

3.7. Многоязычный поиск web-документов (WebCLEF)

Коллекция документов для данной дорожки собрана с web-страниц правительственных сайтов европейских стран [5, 6]. Общий объем коллекции — ~3.5 миллиона страниц. Коллекция включает документы на 20 языках с 27 доменов первого уровня.

Список заданий для данной дорожки состоял из 1940 запросов, часть из которых была сформирована автоматически.

3.8. Многоязычный поиск с учетом геграфических отношений (GeoCLEF)

В данной дорожке участникам предлагались задачи многоязычного поиска документов (Ad Hoc), при этом запросы включали различные географические отношения [7]. Например:

- Города не далее 100 километров от Франкфурта, Германия.
- Винодельческие регионы около рек Европы.
- Малярия в тропиках.

Анализ географических отношений не является обязательным для участия в данной дорожке. Таким образом, GeoCLEF позволяет сравнить эффективность стандартных методов ad-hoc поиска (с учетом, естественно, многоязычия) и специализированных методов, включающих анализ географических отношений.

4. Некоторые личные наблюдения

Многоязычие

Несмотря на очевидный рост интереса научного сообщества к задачам многоязычного поиска (см. рис. 1 выше), по-видимому, отражающий и определенную общественную потребность, наблюдается некоторый дефицит реальных приложений многоязычного поиска, и дефицит готовых к применению систем многоязычного поиска. Действительно, немногим пользователям актуально находить документы на незнакомом языке при отсутствии адекватной системы перевода или кросс-языкового анализа найденных документов.

Однако, есть и успешные примеры применения кросс-языкового поиска. Например, библиотека англо- и немецко-язычных документов по социальным наукам предназначена в основном для ученых из Германии, которые хорошо понимают английский язык, но с удобством используют систему, которая позволяет ввести запрос на одном языке, а получить документы на обоих языках.

Сообщество CLEF постоянно ищет актуальные задачи многоязыкового поиска, которые могут быть востребованы большим количеством людей. В частности, в 2006 году появились интересные задача поиска по Википедии, Flickr (см. выше).

В самом деле, содержание фотографии (Flickr) может быть понятно человеку, даже если комментарий к фотографии написан на незнакомом языке. А в случае онлайновой многоязычной энциклопедии (Википедии), было бы полезно объединить усилия авторов из разных стран по сбору информации, выявлять противоречия в описаниях статей на разных языках.

Формулировка потребности пользователя

Постановка задачи в дорожке ad-hoc и подобных (Domain-Specific, WebCLEF) такова: найти все документы, содержащие полезную информацию по данному топику. Формулировка топика состоит из трёх частей: Title, Description, Narrative.

Например:

```
<top>
<num> 148 </num>
<EN-title> Russian Germans and their Language </EN-title>
<EN-desc> Find documents that discuss the linguistic
   integrity of Russian Germans from the former Soviet Union
   and/or Russia now living in Germany or Russia. </EN-desc>
<EN-narr> Relevant documents report on Russian Germans and
   the language that they currently speak or use. Documents
   are relevant if they explicitly discuss the practising or
   learning of the German or Russian languages. Documents may
   discuss linguistic considerations from a cultural or
   identity-forming view as well as language-related aspects
   of the integration of emigrants or non-emigrants.
   Documents not explicitly focussing on language as a
   central topic are irrelevant. </EN-narr>
</top>
```

Информационная система может использовать любые из указанных трех полей, но обычно наиболее эффективные результаты получаются при использовании только Topic, либо Topic+Description.

Таким образом, по сравнению с РОМИП ad-hoc, запросы получаются в среднем более длинными и внешне более «естественно-языковыми».

Казалось бы, длинные естественно-языковые запросы позволяют использовать более глубокий анализ языка и сложные алгоритмы? Однако, большинство систем используют различные вариации bag-of-words алгоритмов[1].

Значительные усилия прилагаются для улучшения переводов с языка на язык — тут применяются сложные комбинации систем машинного переводы и статистической многозначной трансляции [8].

Также популярны техника pseudo-relevance feedback, а вот учет расстояний между словами — не популярен.

Сравнительная эффективность кросс-языкового поиска

Для оценки эффективности кросс-языкового поиска применяется сравнение с basic line — применением тех же алгоритмов для моноязычного поиска.

Обычно средние значения метрик качества кросс-языкового поиска ниже соответствующих средних значений метрик для

[1] В частности, в одном из докладов сообщалось, что статистический анализ текстов публикаций CLEF показал, что наиболее часто употребляемый термин — BM25.

моноязычного поиска. Однако, анализ по отдельным заданиям показывает, что для некоторых заданий эффективность кросс-языкового поиска оказывается существенно выше, чем эффективность соответствующего моноязычного алгоритма.

Заключение

Развитие CLEF показывает ведущие и перспективные направления развития информационного поиска как научного направления и, в частности, многоязычного информационного поиска. В частности, отметим:

- непрерывно возрастающий интерес научного сообщества к проблемам информационного поиска;
- смещение фокуса внимания от классического ad-hoc поиска к современным задачам мультимедийного, географического поиска, поиска точных ответов на заданный вопрос;
- новые, современные задачи требуют применения широкого арсенала методов анализа текстов и данных.

Задачи интеграции многоязычных коллекций особенно актуальны для Европейского Союза. Вместе с тем, существуют задачи многоязычного поиска, которые могут быть востребованы российскими пользователями, например:

- рассматриваемые в рамках CLEF задачи — поиск в Википедии, Flickr, других мультимедийных коллекциях;
- поиск правовой информации на различных языках — для взаимодействия с иностранными партнерами;
- поиск патентной, научной информации в многоязыковых коллекциях.

Литература

[1] Cross-Language Evaluation Forum (CLEF) http://www.clef-campaign.org

[2] Carol Peters. What happened in CLEF 2006 Introduction to the Working Notes. // Proceedings of CLEF'2006 http://www.clef-campaign.org/2006/working_notes/workingnotes2006/petersCLEF2006.pdf

[3] Bernardo Magnini et. al. Overview of the CLEF 2006 Multilingual Question Answering Track // Proceedings of CLEF'2006 http://www.clef-campaign.org/2006/working_notes/workingnotes2006/magniniOCLEF2006.pdf

[4] Douglas W. Oard et. al. Overview of the CLEF-2006 Cross-Language Speech Retrieval Track // Proceedings of CLEF'2006

http://www.clef-campaign.org/2006/working_notes/workingnotes2006/oardOCLEF2006.pdf
[5] Bцrkur Sigurbjцrnsson, Jaap Kamps, Maarten de Rijke. Overview of WebCLEF 2005 // Proceedings of CLEF'2005. LNCS 4022, Springer — 2006 — pp. 810-824.
http://www.clef-campaign.org/2005/working_notes/workingnotes2005/sigurbjornsson05.pdf
[6] Krisztian Balog, Leif Azzopardi, Jaap Kamps, Maarten de Rijke. Overview of WebCLEF 2006 // Proceedings of CLEF'2006
http://www.clef-campaign.org/2006/working_notes/workingnotes2006/balogOCLEF2006.pdf
[7] Fredric Gey et. al. GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview // Proceedings of CLEF'2006
http://www.clef-campaign.org/2006/working_notes/workingnotes2006/geyOCLEF2006.pdf
[8] Atelach Alemu Argaw, Lars Asker. Amharic-English Information Retrieval.
http://www.clef-campaign.org/2006/working_notes/workingnotes2006/atelachalemuargawCLEF2006.pdf
[9] M. Ageev, B. Dobrov, N. Loukachevitch. Sociopolitical Thesaurus in Concept-based Information Retrieval // Proceedings of CLEF'2005. LNCS 4022, Springer — 2006 — pp. 141-150.
http://www.clef-campaign.org/2005/working_notes/workingnotes2005/sigurbjornsson05.pdf
[10] Агеев М.С., Кураленок И.Е. Официальные метрики РОМИП'2004 // Российский семинар по Оценке Методов Информационного Поиска (РОМИП 2004) — Пущино, 2004.