LEGAL NOTICE:

Copyright for individual papers in these proceedings resides with the individual authors of each paper. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. SIGIR Workshop on Information Access in a Multilingual World July 23, 2009 Boston, Massachusetts USA

Program Committee:

Martin Braschler, Zürich University of Applied Sciences, Switzerland
Aitao Chen, Yahoo! Research, USA
Kuang-hua Chen, National Taiwan University, Taiwan ROC
Ruihua Chen, Microsoft Research, China
Nicola Ferro, University of Padua, Italy
Atsushi Fujii, Tsukuba University, Japan
Julio Gonzalo, Universidad Nacional de Educación a Distancia, Spain
Gareth Jones, Dublin City University, Ireland
Kazuaki Kishida, Keio University, Japan
Sadao Kurohashi, Kyoto University, Japan
Kazuko Kuriyama, Shirayuri College, Japan
Gina-Anne Levow, University of Chicago, USA
Chin-Yew Lin, Microsoft Research, China
Thomas Mandl, University of Hildesheim, Germany
James Mayfield, Johns Hopkins University, USA
Mandar Mitra, Indian Statistical Institute, India
Tatsunori Mori, Yokohama National University, Japan
Isabelle Moulinier, Thomson Reuters, USA
Paul McNamee, Johns Hopkins University, USA
Douglas W. Oard, University of Maryland, USA
Maarten de Rijke, University of Amsterdam, The Netherlands
Miguel Ruiz, University of North Texas, USA
Yohei Seki, Toyohashi University of Technology, Japan
Benjamin Tsou, City University of Hong Kong, China
Takehito Utsuro, Tsukuba University, Japan
Christa Womser-Hacker, University of Hildesheim, Germany
Kam-Fai Wong, Chinese University of Hong Kong, China
Masaharu Yoshioka, Hokkaido University, Japan
Introduction and Overview

This workshop is the third workshop on the topic of multilingual information access held during SIGIR conferences this decade. The first was at SIGIR 2002 on the topic of “Cross Language Information Retrieval: A Research Roadmap”. The second was at SIGIR 2006 on the topic of “New Directions in Multilingual Information Access”. Over the past decade the field has matured and real world applications have appeared. Thus our goal in this 2009 workshop was to collate experiences and plans for the real-world application of multilingual technology to information access. Our aim is to identify the remaining barriers to practical multilingual information access, both technological and from the point of view of user interaction. We were fortunate to obtain as invited keynote speaker Dr Ralf Steinberger of the Joint Research Centre of the European Commission, presenting the Joint Research Centre's multilingual media monitoring and analysis applications, including NewsExplorer. Dr. Steinberger has written an overview paper about their family of applications, which is the first paper in these proceedings.

In our call for papers we specified two types of papers, research papers and position papers. Of the 15 papers initially submitted, two were withdrawn and two were rejected. We accepted 3 research papers and 8 position papers, covering topics from evaluation (of image indexing and of cross-language information retrieval in general), Wikipedia and trust, news site characterization, multilinguality in digital libraries, multilingual user interface design, access to less commonly taught languages (e.g. Indian subcontinent languages), implementation and application to health care. We believe these papers represent a cross-section of the work remaining to be done in moving toward full information access in a multilingual world.

We would like to thank our program committee who read, reviewed and wrote reviews under major time constraints (in some cases 2 days).

Fredric Gey, Noriko Kando and Jussi Karlgren
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Paper title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>An introduction to the Europe Media Monitor: family of applications ..........</td>
<td>1</td>
</tr>
<tr>
<td>Ralf Steinberger, Bruno Pouliquen, Erik van der Goot (European Commission Joint Research Centre)</td>
<td></td>
</tr>
<tr>
<td>The Assignment of Tags to Images in Internet: Language Skill Evaluation for Tag Recommendation</td>
<td>9</td>
</tr>
<tr>
<td>David Nettleton, Mari-Carmen Marcos (Pompeu Fabra University), Bartolomé Mesa-Lao (Autonomous University of Barcelona)</td>
<td></td>
</tr>
<tr>
<td>Multilingual Wikipedia, Summarization, and Information Trustworthiness ...</td>
<td>19</td>
</tr>
<tr>
<td>Elena Filatova (Fordham University)</td>
<td></td>
</tr>
<tr>
<td>NSContrast: An Exploratory News Article Analysis System that Characterizes the Differences between News Sites</td>
<td>25</td>
</tr>
<tr>
<td>Masaharu Yoshioka (Hokkaido University)</td>
<td></td>
</tr>
<tr>
<td>What’s wrong with Cross-Lingual IR? ..................................................</td>
<td>31</td>
</tr>
<tr>
<td>John I. Tait (Information Retrieval Facility)</td>
<td></td>
</tr>
<tr>
<td>A Virtual Evaluation Track for Cross Language Link Discovery .................</td>
<td>33</td>
</tr>
<tr>
<td>Wei Che (Darren) Huang (Queensland University of Technology), Andrew Trotman (University of Otago), Shlomo Geva (Queensland University of Technology)</td>
<td></td>
</tr>
<tr>
<td>Towards an Integrative Approach to Cross-Language Information Access for Digital Libraries</td>
<td>41</td>
</tr>
<tr>
<td>Jiangping Chen, Miguel E Ruiz (University of North Texas)</td>
<td></td>
</tr>
<tr>
<td>Ubiquity: Designing a Multilingual Natural Language Interface .................</td>
<td>45</td>
</tr>
<tr>
<td>Michael Yoshitaka Erlewine (Mozilla Labs)</td>
<td></td>
</tr>
<tr>
<td>Romanization – An Untapped Resource for Out-of-Vocabulary MT for CLIR ..</td>
<td>49</td>
</tr>
<tr>
<td>Fredric Gey (University of California, Berkeley)</td>
<td></td>
</tr>
<tr>
<td>Urdu is not Hindi for Information Access ..........................................</td>
<td>53</td>
</tr>
<tr>
<td>Kashif Riaz (University of Minnesota)</td>
<td></td>
</tr>
<tr>
<td>A Patient Support System based on Crosslingual IR and Semi-supervised Learning ..................................................</td>
<td>59</td>
</tr>
<tr>
<td>Hideki Isozaki, Tsutomu Hirao, Katsuhito Sudoh, Jun Suzuki, Akinori Fujino, Hajime Tsukada, Masaaki Nagata (NTT Corporation)</td>
<td></td>
</tr>
</tbody>
</table>
An introduction to the Europe Media Monitor family of applications

Ralf Steinberger, Bruno Pouliquen & Erik van der Goot
European Commission
Joint Research Centre (JRC)
21027 Ispra (VA), Italy
Tel: (+39) 0332 78 5648
Ralf.Steinberger@jrc.ec.europa.eu

ABSTRACT
Most large organizations have dedicated departments that monitor the media to keep up-to-date with relevant developments and to keep an eye on how they are represented in the news. Part of this media monitoring work can be automated. In the European Union with its 23 official languages, it is particularly important to cover media reports in many languages in order to capture the complementary news content published in the different countries. It is also important to be able to access the news content across languages and to merge the extracted information. We present here the four publicly accessible systems of the Europe Media Monitor (EMM) family of applications, which cover between 19 and 50 languages (see http://press.jrc.it/overview.html). We give an overview of their functionality and discuss some of the implications of the fact that they cover quite so many languages. We discuss design issues necessary to be able to achieve this high multilinguality, as well as the benefits of this multilinguality.

Categories and Subject Descriptors
H.3.1 [Content Analysis and Indexing]: Linguistic Processing, Thesauruses; H.3.3 [Information Search and Retrieval]: Clustering, Information Filtering; H.3.5 [Online Information Services]: Web-based Services;

General Terms
Languages, Algorithms, Design, Theory.

Keywords
Media monitoring; multilinguality; cross-lingual information access; CLIA; information extraction; Europe Media Monitor.

1. INTRODUCTION
Most large organizations have dedicated departments that monitor the media to keep up-to-date with relevant developments and to keep an eye on how they are represented in the news. Specialist organizations such as those that monitor threats to Public Health, monitor the multilingual media continuously for early warning and information gathering purposes. In Europe and large parts of the world, these organizations look at the news in various languages because the content found across different languages is complementary. In the context of the European Union, which has 23 official languages, multilinguality is a practical and a diplomatic necessity, and so is the related ability to access information across languages.

Current approaches to link and access related textual content across languages use Machine Translation, bilingual dictionaries, or bilingual vector space representations. These approaches are limited to covering a small number of languages. We present an alternative way of tackling this challenge. It consists of aiming at a language-neutral representation of the multilingual document contents.

The European Commission’s Joint Research Centre (JRC) has developed a number of multilingual news monitoring and analysis applications, four of which are publicly accessible to the wider public. All four systems can be accessed through the common entrance portal http://press.jrc.it/overview.html.

Users of the systems are the EU Institutions, national organizations in the EU Member States (and in some non-EU countries), international organizations, as well as the public. The freely accessible websites receive between one and two million hits per day from thirty to fifty thousand distinct users.

The next section provides information on the news data that is being analyzed every day. Section 3 gives an overview of the four publicly accessible media monitoring systems and their specific functionalities. Section 4 aims to explain the specific approach which enabled the EMM family of applications to cover a high number of languages and to fuse information extracted from texts written in different languages. It also describes some of the multilingual and cross-lingual functionalities that we could only develop because we adopted this approach. The last section points to future work. Related work for each of the functionalities described will be discussed in the related sections.

2. The Europe Media Monitor news data
The Europe Media Monitor (EMM, see [1]) is the basic engine that gathers a daily average of 80 to 100,000 news articles in approximately 50 languages (status June 2009), from about 2,200 hand-selected web news sources, from a couple of hundred...
specialist and government websites, as well as from about twenty commercial news providers. EMM visits the news web sites up to every five minutes to search for the latest articles. When news sites offer RSS feeds, EMM makes use of these, otherwise it extracts the news text from the often complex HTML pages. All news items are converted to Unicode. They are processed in a pipeline structure, where each module adds additional information. Whenever files are written, the system uses UTF-8-encoded RSS format.

3. The Europe Media Monitor applications

The EMM news gathering engine feeds its articles into the four fully-automatic public news analysis systems, and to their non-public sister applications. The major concern of NewsBrief and MedISys (see [18]) is breaking news and short-term trend detection, early alerting and up-to-date category-specific news display (Sub-section 3.1). NewsExplorer (3.2) focuses on daily overviews, long-term trends, in-depth analysis and extraction of information about people and organizations. EMM-Labs (3.3) is a collection of more recent developments and includes various tools to visualize the extracted news data.

For NewsBrief and MedISys, there are different access levels, distinguishing the entirely public web sites from an EC-internal website. The public websites do not contain commercial sources and may have slightly reduced functionality.

3.1 Live news monitoring and breaking news detection in NewsBrief and MedISys

All EMM news items get fed into NewsBrief and into the Medical Information System MedISys as soon as they come in. While NewsBrief is a wide-coverage monitoring tool covering the interests of all users, MedISys sieves out those news reports that talk about potential Public Health threats, including those of chemical, biological, radiological and nuclear origin (CBRN). Every ten minutes and in each of the languages, both applications cluster the latest news items (four hour window or more, depending on the number of recent articles) and present the largest cluster as the current top-ranking media theme (Top Stories). The title of the cluster’s medoid (the article closest to the cluster centroid) is selected as the most representative title and thus as the title for the cluster. All current clusters are compared to all the clusters produced in the previous round. If at least 10% of the articles overlap between a new cluster and any of the previous ones, the clusters get linked and those articles that have fallen out of the current 4-hour-window get attached to the current cluster. The public web pages are updated every the minutes so that users always see the latest news of the fastest news providers.

Larger new clusters (without overlap to previous clusters) and clusters of a rapidly rising size get automatically classified as breaking news so that subscribed users will be notified by email. The statistical breaking news detection algorithm makes use of information on the number of articles and of the number of different news sources, comparing the news of the last 30 minutes with longer periods of time.

Each article is geo-tagged, i.e. potential place names are identified and ambiguities are resolved (for a description of such ambiguities, see Section 0). An algorithm that considers the place hierarchy (city is part of a region which is part of a country) and the frequency of mentions determines the major location in a cluster. This is used to visualize the location of the current news items on a geographical map (see Figure 1).

All news items are additionally categorized into hundreds of categories. Categories include geographic regions such as each country of the world, organizations, themes such as natural disasters or security, and more specific classes such as earthquake, terrorism or tuberculosis. Articles fall into a given category if they satisfy the category definition, which consists of Boolean operators with optional vicinity operators and wild cards. Alternatively, cumulative positive or negative weights and a threshold can be used. Uppercase letters in the category definition only match uppercase words, while lowercase words in the definition match both uppercase and lowercase words. Many categories are defined with input from the institutional users themselves.

The system keeps statistics on the 14-day average number of articles falling into any given country-category combination (e.g. Poland-tuberculosis). If the number of articles for this combination found in the last 24-hours (normalized by weekday fluctuations) is significantly higher than this average, a country-category-specific alert is triggered and users are notified using alert ranking graphs (see Figure 2).

As the categories are defined in many different languages (depending on the user interests, some are defined in all languages, others in only a few), the statistics are fed by all languages and are thus not language-dependent. This means that the sensitive alerting tool will detect a sudden rise in any of the languages, meaning that users may see an alert even before the event is reported in their own language. For humanitarian and
public health institutions whose main concern is early warning and rapid reaction, this is a highly appreciated functionality.

3.2 Trend monitoring and information extraction in NewsExplorer

NewsExplorer takes as input the non-commercial EMM articles in 19 languages collected within a calendar day, clusters them separately for each of its languages and displays the major news per day, ordered by cluster size. Languages covered include 14 EU languages, and additionally Arabic, Farsi, Russian, Turkish and Norwegian. NewsExplorer applies a number of text mining tools to detect names of persons, organizations and locations and displays those names found in the course of each day. For each cluster, a dedicated web page gets created, where the same type of information is displayed specifically for this cluster. It furthermore detects reported speech quotations by people (X said “…”) and about people (Y said “…X…”), as well as titles of persons used in the media (e.g. former foreign minister, playboy, 58-year-old). Many persons are referred to using variant spellings, not only across writing systems (e.g. Greek Γκέρχαρν Σρέντερ, Russian Герхард Шредер, and German Gerhard Schröder all refer to the former German Chancellor) and across languages (e.g. English Vladimir Ustinov and German Wladimir Ustinow), but also within the same language. For instance, the following variants were found within English language news:

Schröder, Schroeder, Schröeder. NewsExplorer aims to detect these variants as belonging to the same person. It additionally retrieves more name variants (e.g. Chinese, Japanese and Thai) from Wikipedia, resulting in up to 170 name variants for the same person. Figure 3 shows some name variants for the current US President. Users searching the system for any of these will also find all articles using any of the alternative spellings. As of June 2009, the name database contains 900,000 names plus about 170,000 variants. Names from the name database are exported daily into a finite-state automaton so that these known names will also be found in EMM’s live news monitoring systems described in Section 3.1.

For each person or organization name (referred to as entity) detected in at least two different news articles in the same cluster, NewsExplorer displays all the extracted meta-information on individual news pages, one per entity. NewsExplorer also computes which entities get frequently mentioned together (displayed as related names) and – more interestingly – which entities get mentioned frequently with the first person but do not have much media attention for themselves, using a TF.IDF-like formula (see [15]). This latter group of persons, referred to as associated names, typically consists of close contacts such as relatives, personal spokespersons or secretaries, etc. An interactive visualization tool allows users to display entities, associated names, and names jointly associated to various entities (See Figure 4).

For all the daily clusters, NewsExplorer establishes and displays links to related clusters found in the previous days. Many events or themes capture the media’s attention for a longer period, such as the Israel-Palestine or the Iran-nuclear conflicts. NewsExplorer links all these historically related news clusters into stories. For each story, an interactive graph visually presents media reporting activity over time (Figure 5) and displays story-specific meta-information, such as names, countries and keywords extracted from all the – sometimes thousands – related news articles. Together with the name variant mapping tool (see also 4.3.3), the probably most interesting feature of NewsExplorer is the cross-lingual cluster linking functionality: for each cluster, users can – with a simple click on the mouse – jump straight to the equivalent news reports in the other languages. NewsExplorer establishes links between clusters in different languages if the multilingual clusters have a minimum cross-lingual similarity based on subject
domains, person and organization names, locations and cognates (words that are the same across languages). This application is described in Section 4.3.1.

3.3 Further visualization and analysis tools in EMM-Labs
EMM-Labs is a loose collection of further media-based text mining and interactive visualization tools. It includes generic or topic-related geographical news maps (e.g. on swine flu), various types of statistics on EMM categories over time and across languages, and social network browsers showing generic co-occurrence frequency relations between persons or specific extracted relations such as support, criticism and family relations.

The news cluster navigation tool allows an interesting alternative view of the news: it graphically shows an interactive network of the latest news clusters, the most active categories at the moment, and the countries most mentioned in the news right now, with stronger and weaker links between the nodes belonging to any of the three groups (see Figure 6). The interactive interface of the tool lets users decide on various thresholds (numbers of clusters displayed, percentage of articles of a cluster that need to be classified according to country or category for the link to be displayed, etc.).

The event extraction tool detects violent events, accidents, natural and humanitarian disasters and more in the live news clusters (fed by NewsBrief) in the languages English, Russian, French, Italian and Spanish, as well as in Arabic (after an analysis of the English text machine-translated from Arabic). The tool detects and displays information on the event type, on the location of the event, as well as on number and status of the victims (see Figure 7 and Figure 8). The displayed facts are a best-guess combination of the information found in all the articles of the news cluster. For details on this application, see [2].

4. Approach to multilingual and cross-lingual information processing
Since their conception, the prerequisite for EMM applications was that they should be highly multilingual, aiming at all EU languages and possibly more. This challenging objective was to be achieved with a relatively small-sized developer team, so that it was clear that any components and processes needed to be simple and, ideally, language-independent. While – to some extent – the simplicity was a hindrance as it did not allow to consider many language-specific features, the achieved multilinguality brought many advantages not usually achieved by systems covering one or only a few languages. In this section, we want to highlight some of the methods used and describe some of the benefits of multilinguality in EMM.
languages
tuberculosis
various languages (e.g. word '
pain'
things across language) may sometimes lead to false positives.
disadvantage that false friends (same words meaning different
e.g. definition words are all mixed inside the same definition and
all languages combined, meaning that English, French, Russian,
many languages. EMM uses only one definition per category for
Category definitions in
NewsBrief and in MedISys are
multilingual. News items from different languages thus get
categorized into exactly the same classes. While we use the
English name as the class label, the classes themselves are thus
being filled with news articles from many different languages.
This has the advantage that trend graphs showing an increase in
articles per category may alert users about increased activity in a
specific field even before news articles in their own language are
published (see Figure 2 and the explanations in Section 3.1). The
categorization of texts according to the countries mentioned is
treated in exactly the same way, i.e. there is one EMM class for
each country of the world. This allows to display multilingual
country-category information (the intersection of any country
category with any subject domain category) on maps (see Figure
9). Note that the geo-tagging application described in Section 4.2
is performed independently of this country categorization.

The method to categorize the articles, described in Section 3.1, is
rather simple and user-friendly, and it lends itself to dealing with
many languages. EMM uses only one definition per category for
all languages combined, meaning that English, French, Russian,
etc. definition words are all mixed inside the same definition and
are thus all searched in all languages. This has the obvious
disadvantage that false friends (same words meaning different
things across language) may sometimes lead to false positives.
For instance, users searching for the disease symptom ‘pain’ may
erroneously receive French articles about bread as the French
word ‘pain’ means bread. However, these multilingual definition
lists also have advantages: (a) They are very easy to maintain –
which is not a negligible criterion when dealing with hundreds of
categories in about 50 languages; (b) they automatically benefit
from cognate words that are the same across many languages (e.g. tsunamis), or that are almost the same, so that one simple search
word with wildcards will often allow to match the words in
various languages (e.g. tuberculosis, German Tuberkulose, French tuberculeux and
tuberculose, etc.). The obvious alternative to producing hand-
crafted category definitions would theoretically be the usage of
machine learning methods for categorization (e.g. using Support
Vector Machines or Naïve Bayes), but most EMM users would
not be willing to provide training examples in all the languages.
For many categories, such as diseases, organizations or subjects
(such as genetically modified organisms, GMO) it is much easier
to gather multilingual search terms from Wikipedia or to ask
colleagues for help who speak the languages of interest. Users
regularly request the monitoring of new news categories, and due
to the simplicity of the approach, first versions of the category
definitions can often be produced within a day.

4.2 Multilingual information extraction using
mostly language-independent rules
Developing text mining components such as tools to extract
structured information from unstructured free-text typically
requires writing (grammar) rules and using dictionaries.
Developing such tools for many languages is rather labor-
intensive. Developing time can be saved by re-using such
linguistic resources of one language to develop the resources for
the other languages (see, e.g. [4], [10], [9]). However, even
adapting existing resources to new languages is still rather time-
consuming. In NewsExplorer, which carries out Information
Extraction (IE) for 19 languages, we thus aimed at using the
same, language-independent rules for all languages, and to store
any language-dependent information in language-specific
parameter files. From the conception phase of any text analysis
tool development in NewsExplorer, the multilinguality
requirement is an unmovable premise. For details, and for
examples how these principles were realized in eight different text
analysis applications, see [20].

A simplified example of such a language-independent rule is the
quotations recognition pattern below, which extracts reported
speech and the person issuing the quotation, by making use of the
slots name, reporting-verb and quote-mark. The set of possible
quote marks (such as <<, >>, “”, etc.) is mostly the same across
languages. The English language-specific parameter file includes
lists of reporting verbs such as said, reported, added, etc. and
their morphological variants (e.g. says, was saying). Person names
are identified and marked up using a separate tool, which is based
on the same principles. Further slots (not shown here) may
include modifier and auxiliary-verb for such a rule. Elements
within square brackets are optional.

name [:|up to 60 chars|] reporting-verb [:|that|] quote-mark
QUOTE quote-mark

e.g. John Smith, supporting AFG, said: "They are the best!".

Another example for language-independent rules is the
geo-tagging application, which includes geo-parsing (identification of
potential geographic references in text) and disambiguation. Geo-
tagging requires a gazetteer (containing lists of locations and their
geo-co-ordinates), which is thus an unavoidable linguistic
resource. Disambiguation is necessary due to various types of
homography, i.e. between locations and persons (Paris in France
vs. Paris Hilton), between locations and generic words (e.g. Split
is a city in Croatia) and between locations which share the same
name (e.g. there are 15 locations world-wide called Paris and
there are 32 places called Washington). The geo-disambiguation
rules in NewsExplorer are language-independent, as they make
use of features such as the information whether an entity has
previously been tagged as being a person or organization name,
gazetteer-provided information on the size class of the locations
over one hundred language pairs

Examples for cross-lingual information access applications are cross-lingual information retrieval (CLIR), cross-lingual information access (CLIA, e.g. cross-lingual glossing, or the linking of related documents across languages), cross-lingual name variant matching (identifying that Ali Chamenei, Ali Jumenei and Asu Xaneou are the same name), cross-lingual plagiarism detection, multilingual summarization and machine translation (MT). Known approaches to these applications make use of MT (e.g. [8]) or of different kinds of bilingual dictionaries or thesauri (manually produced or machine-generated, e.g. [24]). An alternative approach consists of producing bilingual word associations by feeding pieces of parallel text to systems applying Lexical Semantic Analysis (LSA, [6]) or Kernel Canonical Correlation Analysis (KCCA, [23]). All of these approaches have in common that they are basically bilingual ([3] being a notable exception), meaning that multilingual applications are actually bilingual. In a highly multilingual environment with N languages, there are vast numbers of language pairs, namely: \( N^2 - N - 1 \). In NewsExplorer, which covers 19 languages, there are thus 171 language pair combinations, and 342 language pair directions. A multi-bilingual approach clearly is impractical.

### 4.3 Components used in NewsExplorer

For NewsExplorer’s components that link related documents across languages or that fuse information found in different languages, the only possible solution was thus to work with an interlingua type of approach, i.e. representing contents in a language-neutral way and mapping contents in each of the languages (news articles, news clusters, or extracted pieces of information) to this language-independent representation. The following representations are used in NewsExplorer (for details, see [14]):

(a) A vector of multilingual subject domain representations using the Eurovoc thesaurus (see [16] and Section 4.3.2);
(b) a frequency list of locations per document, represented by location identifiers and their latitude-longitude information (see [13] and Section 4.3.3);
(c) a frequency list of person and organization name identifiers (also described in Section 4.3.3);
(d) monolingual weighted lists of keywords. Identifiers (b) and (c) each represent a whole range of multilingual variants. Sections 4.3.2 and 4.3.3 give details on these four ingredients for cross-lingual cluster linking.

Further possible alternatives for a language-neutral representation (not currently exploited), are normalized expressions to express dates or measurement units (speed, acceleration, time, etc.), subject-specific multilingual nomenclatures such as the Medical Subject Headings MeSH or the Customs Tariff Code TARIC, and more (see [19] for further ideas and more detail).

### 4.3.2 Multilingual subject domain representation

The idea of ingredient (a) in 4.3.1 was thus to represent the contents of documents by a ranked list of subject domains (also referred to as topic maps), using a multilingual classification of subject domains that are the same across languages. The only resource we are aware of that is available to be used for such a purpose is Eurovoc. Eurovoc is a wide-coverage thesaurus with over 6,000 classes that was developed for human subject domain classification in the European Commission and in parliaments in the European Union and beyond. As Eurovoc has been used for many years, tens of thousands of documents exist that have been labeled manually by librarians. The documents in 22 languages which are part of the freely available multilingual parallel corpus JRC-Acquis ([21]) are accompanied by their respective Eurovoc codes. The classifier that assigns Eurovoc codes to news clusters in NewsExplorer was developed by applying machine learning techniques, using these manually classified documents. The major challenges of this task were that this is a multi-label task (each manually classified document belongs on average to 5.6 classes) involving a large number of categories (more than 3,000 of the 6,000 Eurovoc classes are actively used) with a highly imbalanced category distribution and heterogeneous text types. The adopted approach, described in [16], produces a ranked and weighted list of (numerical) Eurovoc subject domain codes. In NewsExplorer, the cosine similarity is applied to compare this language-independent subject domain representation for the daily news clusters in one language with those in other languages. This is the major ingredient for the cross-lingual cluster similarity measure in NewsExplorer.

### 4.3.3 Language-neutral entity vectors

The second most important ingredient for cross-lingual cluster linking in NewsExplorer ((b) in 4.3.1) is the similarity based on geo-references. We described in Section 4.2 the principles behind our geo-tagging application, which recognizes potential location names in free text and resolves any ambiguities using language-independent rules. This tool produces, for each document in each of the languages, a frequency list of locations including the country they belong to. At cluster level, NewsExplorer aggregates this information by counting the references to each country (e.g. one mention of the Turkish city of Izmir will add one count to the country count of Turkey, which is itself represented by the language-neutral country ISO code) and produces a frequency count of (direct or indirect) country mentions. The cosine similarity is applied to the country vector representation of each pair of clusters in the different languages to calculate the value of the second ingredient.

The third ingredient vector (c) in 4.3.1 consists of a frequency list of normalized person or organization names, where the multilingual name variants for the same entity are all represented by one language-neutral numerical code (see Figure 3). Named-entity recognition rules are applied to all the news in the NewsExplorer languages. The rules were developed according to

---

the principles laid out in Section 3.2. These rules identify name mentions in each of the languages. If the name variants found are already known, the names will be represented by the language-neutral entity identifier. If the name mention found is a previously unknown name, a name variant matching procedure will be launched. This procedure first transliterates non-Latin script names into the Latin character set, using standard hand-crafted transliteration rules. It then normalizes all names by applying about 30 different rules, which try to map language-specific spelling differences onto one canonical form. Examples for such empirically found differences are, for instance, that diacritics of foreign names are often omitted (e.g. outside Poland, the name Lech Wałęsa is usually spelt as Wałęsa) and that consonants are frequently doubled or singled (e.g. Mohammed vs. Mohamed and Barack vs. Barrak, see Figure 3). Russian names ending in -ov (-ov) may furthermore be spelled as -ев, -оев, -ев, etc., depending on the target language of the transliteration. After applying these normalization steps, which aim at mapping frequent spelling variations to one canonical spelling, all vowels are removed because vowels often differ, especially for transcriptions of Arabic names. The result of the name normalization is the canonical form of this name (e.g. mohamed tayyib for Mahmoud Ahmadinejad, and kondza rc for Condoleezza Rice). The canonical form of each newly found name is then compared to the canonical form of all known names and their variants. If the two canonical forms are the same, the Levenshtein edit distance is applied to two different representations of the name: (1) the found surface string of a name or the transliteration result, and (2) the normalized form with vowels. If the combined similarity between two names is above an empirically established threshold, the name variants get automatically merged and the new name becomes a known name variant of an existing name. In the course of five years of daily analysis, NewsExplorer has altogether identified over one million names and name variants, with as many as 170 different name variants for the same person. For a detailed description of this name variant matching approach, see [12]. Unlike other known methods to map names across scripts ([7], [5]) or to compute string similarity measures ([11]), the presented algorithm is not multi-bilingual, but uses a single representation for all languages. New languages can either be added without any additional steps, or one or more new normalization rules can be added to the set of normalization rules. All normalization rules are applied to all languages. For the purposes of creating the third ingredient for the cross-lingual cluster similarity calculation, each news cluster is represented by a frequency list of all known or new names, with all name variants being represented by the same numerical entity identifier. Each daily news cluster will thus again be compared to all the clusters in the other languages, using the cosine similarity measure.

The fourth ingredient for cross-lingual cluster similarity is simply a log-likelihood-weighted monolingual list of words found in each cluster. The cosine similarity is again applied to compare the vector representations in the different languages. While most of the words are of course different across languages, this fourth similarity still adds to the other three similarities as it considers cognates, names, name parts, numbers and acronyms.

The four ingredients are combined with a relative weight of 0.4, 0.3, 0.2 and 0.1. These weights were set intuitively and after comparing various examples manually, but they have not yet been confirmed empirically. See [14] and [20] for more detail.

4.3.4 Fusion of information about entities found in documents written in different languages

The EMM system’s information extraction components extract different types of information about persons and organizations, including: (a) name variants, (b) name attributes such as titles (e.g. president, stumpman, 58-year-old), (c) news clusters in which the entity was mentioned, (d) stories in which the person was mentioned, (e) quotations by and about the person, (f) co-occurrence information (frequency and a weighted value) for entities being mentioned in the same clusters, (g) labeled relations between persons (English only), namely relations of support, criticism, contact and family relation. EMM applications produce daily and long-term social networks of different types ([17], [22]), based on weighted co-occurrence, based on labeled relations, and based on who mentions whom in reported speech.

As each entity (independently of the name spelling) is represented by one numerical identifier, the information extracted from any of the input languages can be displayed together. To see examples of information gathered across many languages about the same entity, go to NewsExplorer and click on any of the entity names. An additional benefit of having captured the multilingual name variants is that the generated social networks are fed by news articles in many different languages, so that they are independent of the reporting bias by one country or reporting language. Instead, they represent a rather rounded multilingual and multinational picture of the relations between persons.

5. Summary and Future Work

We have given an overview of the functionality of the EMM family of media monitoring applications (Section 3) and we tried to highlight two issues regarding the fact that EMM covers many different languages (Section 4). First, working on many languages while being a small team automatically forced us to using simple means. We tried to show how such simple means could be used to develop relatively complex and powerful applications. Second, we tried to give some examples of a number of applications benefitting from this multilingual news processing.

Currently ongoing work focuses on (a) ensuring a tighter integration of the various applications and tools, (b) adding opinion mining functionality to the existing information extraction components, (c) producing multi-document summaries of the various news clusters in order to allow notifying users more efficiently of breaking news events, and (d) adding blogs as a new text type to the current news monitoring systems. Like for the existing applications, the challenge is to find methods that can be applied to many different languages.

6. ACKNOWLEDGMENTS

The EMM applications have been developed over many years, by many people. We want to thank all contributors for their dedicated work. The following persons should be mentioned explicitly: Flavio Fuart helped to make the EMM output available on robust web pages; Martin Atkinson produced several of the graphical tools, and Jenya Belyaeva contributed with her knowledge of many languages and her thorough quality control.
7. REFERENCES


The Assignment of Tags to Images in Internet
Language Skill Evaluation for Tag Recommendation

David Nettleton  
Department of Information Technology and Communications, Pompeu Fabra University  
Tanger, 122-140  
08018 Barcelona, Spain  
+34 93 542 25 00  
david.netleton@upf.edu

Mari-Carmen Marcos  
Department of Journalism and Audiovisual Communication, Pompeu Fabra University  
Roc Boronat,138  
08018 Barcelona, Spain  
+34 93 542 13 10  
mcarmen.marcos@upf.edu

Bartolomé Mesa-Lao  
Department of Translation and Interpreting, Autonomous University of Barcelona  
Edifici K – Campus UAB  
08193 Barcelona, Spain  
+34 93 581 1876  
barto.mesa@uab.cat

ABSTRACT
Users who tag images in Internet using the English language, can be native in that language or non native. Also, they can have different levels of tagging skills, in semantic terms (richness of vocabulary) and syntactic terms (errors incurred while defining the tags). If we can identify the ‘best’ taggers, we can use their work to help taggers whose skills are not so good. In this paper we present a study carried out for native and non native English language taggers, with the objective of providing user support depending on the detected language skills and characteristics of the user. In order to do this, we study the different types of syntactic errors that taggers commit, and analyze different semantic factors related to objective and subjective tagging, given that our hypothesis is that the latter is in general more difficult. We find that the syntactic and semantic factors together, allow us to profile users in terms of their skill level. This would allow us to keep the tag sessions of the best users and provide their tags to users who have a lower skill level.

Categories and Subject Descriptors
H.3.1 [Content Analysis and Indexing]: Indexing methods.

General Terms
Measurement, Experimentation, Human Factors.

Keywords
Image tagging, tag recommendation, user support, statistical analysis, user study.

1. INTRODUCTION
One of the characteristics of the Web 2.0 is that different web users are responsible for the authorship of the content online and for describing it. This Web, called the ‘social web’, allows any user to find what other users have uploaded, but for this to be possible it is necessary that both (the uploader and the searcher) have in mind similar concepts and words to name (ie, to tag) each object (be it a picture, a video, a web site, a blog, etc.). The uploader describes content and the searcher formulates a query. In this study we assume that tags are influenced by the culture in which each person lives and, of course, by the language they use.

The current general solution to the problem of subjectivity in tags in a free indexing environment is the so-called ‘social indexing’, in which many people index the same object. This gives as a result a set of descriptors which form a ‘folksonomy’, which is intersubjective and therefore, one assumes, a very close approximation to the described object.

English is widely used on the Internet, although for many of the people who use English it is not their native language. In an image tagging context, when non-native English taggers propose tags for an image, due to their limited knowledge of the language, they may define incorrect tags or tags for which there exists a better word. In this paper, we will consider the difficulties for non-native English taggers, how to evaluate their language skill level, and how to offer them appropriate help, such as tag word recommendations and correction.

To this end, we analyze both syntactic and semantic errors in tags in order to detect the degree of skill of a tagger. We consider as syntactic errors: (i) descriptions too long, (ii) spelling errors or typos, (iii) incorrect format (illegal characters), and (iv) incorrect language. As positive semantic factors, we consider: richness (average number of distinct tags used); similarity (frequency of the tags used), and spontaneity (tags used as the first choice). This would classify users according to their syntactic and semantic competence when tagging content. Tags proposed by users with the best competence can be proposed to other users, both native and non-native.

Once confirmed our hypothesis —ie, native speakers have better tagging competence than non-natives—, we proceeded to a...
second phase to determine which type of content is more difficult to tag. For this, we also analyze the differences, by deriving factors, between how users tag objectively (using what we call ‘see’ type tags) and subjectively (by what we call ‘evoke’ type tags). The hypothesis is that ‘evoke’ (subjective) tags require more skill and knowledge of vocabulary than ‘see’ (objective) tags. Therefore, the tagger, and especially the non-native tagger, will require additional help for this type of tags. Image tagging normally does not explicitly differentiate between objective and subjective descriptions. It is assumed that the tagger classifies an image using his/her available vocabulary. But implicitly, the user will probably mix objective and subjective descriptions when tagging an image.

2. RELATED WORK
The so called Web 2.0 or Social Web is one of the most recent phenomenon of Internet which has been studied in recent years [13][4][12][2][11]. In this new web, the users themselves can publish contents produced by themselves and compare them with the rest of the community. Therefore, the Social Web permits any person to search and recover that which other users have published in Internet. Also, in order to make this possible, we need that both users (the user who publishes and the user who retrieves) have in mind similar concepts to name the objects in Internet (web pages, images, videos, etc.) which they wish to share. As a starting point, we assume that the words which are used to represent objects (the labels or tags) will be different depending on the culture of the users and the language they use to describe/recover the contents.

In the literature, studies exist of how a user expresses emotion when interacting with a computer interface [3][7]. In recent years tag recommendation has become an active area of applied research. Different approaches have been made to tag recommendation: based on collective knowledge [14], based on analysis of the images themselves [11][10], collaborative approaches [17][7][8], analysis of folksonomies [9], and systems based on personalization [5]. Although there are not many studies in the literature related to non-native users, we can cite [16] as an example. Finally, one example of a statistical modeling approach can be found in [15].

3. METHODOLOGY – DESIGN OF EXPERIMENTS FOR SKILL EVALUATION
3.1 Selection of the Images
For the present study, we have selected 10 photographs from Flickr. Each image has been selected for what seems to be its ability to require different tags for an ‘objective’ classification, with respect to a ‘subjective’ classification. Image 1 is of a person with his hands to his face; Image 2 is of a man and a woman caressing; Image 3 is of a small spider in the middle of a web; Image 4 is of a group of people dancing in a circle with a sunset in the background; Image 5 is of a lady holding a baby in her arms; Image 6 is of a boy holding a gun; Image 7 is of an old tree in the desert, bent over by the wind; Image 8 is of a hand holding a knife; Image 9 is a photo taken from above of a large cage with a person lying on its floor; finally, Image 10 is of a small bench on a horizon.

3.2 Creation of the Tag Session Web Site
A web site has been prepared which includes a questionnaire which allows the user to introduce his/her demographic data, their tags for the photographs (tag session) and some questions which the user answers after completing the session. The capture of tag sessions was carried out for native and non-native English, and the website can be found at:

3.3 Tag Session Capture
A tag session requires the definition of between 4 and 10 tags which define the objects which are visible in the image. (This number of tags was assigned after evaluating tagging sessions carried out by users in Flickr). The user then defines a new set of tags for the same image, but based on what sensations/emotions the image evokes (for that user). In Figure 1 we can see two columns: in the first column the user defines tags which express what is seen in the image; in the second column the user defines tags which describe what the image evokes. A total of 162 user tag sessions have been captured in the English language, from 2 different countries. For approximately half of the users (from the United States), English is their native language and for the other half (from Spain) it is a second language. The data collection has been carried out between May 2007 and April 2008.

3.4 Data and Factors Derived for Analysis
For each user, we asked them (as part of the questionnaire) for some demographic information: age; gender; nationality; country (where they are tagging); native language. Using this information, together with the tags themselves of the images for each user tagging session, we derived new factors to differentiate the users and tag types.

3.4.1 Summarized and Detailed Descriptive Data about the Tag Session
For each tag session, we collect the following information: language in which the tag session is conducted; easiest image to tag (user is asked); most difficult image to tag (user is asked); the tags themselves assigned for each image, for “See” and “Evoke” separately, and the order in which the tag is assigned. We also record the type of language (if the current tagging language is native or not for the user).

3.4.2 Derived Factors
We define three factors, ‘richness’, ‘similarity’ and ‘spontaneity’, which are calculated and statistically averaged and grouped by user and image:
- Richness: this represents the average number of tags used in “see” and “evoke”. This factor embodies two aspects: (i) images evaluated as easier to tag should have more tags; (ii) users with a wider vocabulary in English should define more tags.
- Similarity: represents the frequency of the tags used for “see” and for “evoke”, relative to other users. We first identify the highest frequency tags for each image (overall for all users), then we compare these tags with those defined by each user to see if the user has used similar tags to everyone else, or if the user has defined more “original” tags. We would assume that a user with a greater vocabulary would be more “original”.

- Spontaneity: this factor compares the first tag a user defines for tag types “see” and for “evoke”, for each image, with the most popular tag (for all users), for each image and tag type.

4. SKILL LEVEL ASSESSMENT

In order to establish the types of support a user needs, we can evaluate the skill level of the user, and the types of errors, if any, that the user commits. We can do this by asking the user to do some tagging examples to calibrate the system, or we can do it ‘on the fly’, while the user is tagging. To identify the user as native or non-native in the tagging language we can: (i) ask him/her at the start of the session; (ii) automatic detection by evaluating the errors made, especially by ‘false friends’ (words adapted from the native language which do not exist in English). This factor is used as the classifier label, and for the present study it is established by method (i) above. We evaluate the skill of the user based on different demonstrated characteristics, which we will now describe. Two levels (or aspects) of the skill level of the tagger are considered: the syntactic level, and the semantic level.

Syntactic level (SYN): evaluated by identifying the following errors: (i) descriptions too long; (ii) spelling errors; (iii) incorrect format (illegal characters); (iv) incorrect language. Depending on the type of error, we can advise the user by appropriate messages.

The percentage of each error type with respect to the total number of tags, gives an approximate syntactic skill level for each user. This skill level should improve as a consequence of the error/guidance messages given to the user.

Semantic level: if we evaluate richness of vocabulary (by the number of different tags they assign for ‘see’ and ‘evoke’), this may correlate with native and non-native. It may also correlate with native taggers who have an inferior vocabulary in their own language.

We also ask the user, during the calibration session, which of the images s/he found the easiest and which were the most difficult to tag. The correlation with the real images which are most difficult/easy (by consensus of the replies of the majority of users), has been incorporated as an additional indication of skill factor (or coherence) which we designate as image evaluation coherence, or ‘IEC’. This factor (IEC) is considered separately from the syntactic and semantic factors. Therefore, we can define a ‘skill level’ formula based on our analysis of the syntactic level, semantic level, and image evaluation coherence (IEC):

Skill level = Semantic_level + IEC – Syn  

where SYN = “avg normalized number of tags with greater than 4 terms” + “average normalized number of spelling errors” + “average normalized number of format errors” + “average normalized number of tags in incorrect language”. Note that we take the average values for the user’s calibration tag session and then we normalize against the worst/highest values found for any user.
**Semantic level** = average normalized number of tags used for 'see' + average normalized number of tags user for 'evoke'. This is the same as the 'richness' derived factor which we described in Section 3.4.2. We then apply a multiplier factor ($\alpha=2$) to semantic level which gives a bias to the part of the skill factor range which is greater than zero.

IEC (Image Evaluation Coherence) = correctly identified the easiest image (0.5) + correctly identified the most difficult image (0.5). Similarly to semantic level, we then apply a multiplier factor ($\beta=2$) to semantic level which gives a bias to the part of the skill factor range which is greater than zero.

Therefore **Semantic level** will have a maximum value of $2 \times 2 = 4$, IEC will have a maximum value of $1 \times 2 = 2$, and **Syntactic level** will have a maximum value of 4.

We observe that the relative weighting of the factors, ($\text{Semantic level, 4}$), ($\text{IEC, 2}$), ($\text{Syntactic level, 4}$) and the multiplier factors $\alpha, \beta=2$, has been calibrated and tested with the real user data. We multiply the ‘positive’ values by $\alpha, \beta=2$ to give a subtotal equal to 6, which gives a slight positive bias to the range of skill level, when we subtract the maximum possible negative value of -4. Therefore, an ‘optimum’ user would have a skill factor of $4 + 2 - 0 = 6$; the ‘worst’ user would have a skill value of $0 + 4 - 4 = 0$; and an ‘average’ user would have a skill value of $2 + 1 - 2 = 1$. Thus, we have derived a factor for ‘skill level’ which can be used to indicate the overall skill level of the user.

5. DATA DEFINITION

In this section we explain how the data captured from the user tag sessions in the Internet questionnaire website, is structured and preprocessed into a format which is adequate for its analysis and modeling. From the questionnaire website, the tagging sessions are loaded into an Excel spreadsheet, and from there they are imported into Microsoft Access. Via a series of preprocesses defined in Visual Basic, the data is reformatted and loaded into the Data Mart. Then, as the final step, the data is processed and loaded into tables which contain the factors and data which are used for the analysis.

5.1 User/Tag Session Datasets

With reference to Tables 1 and 2, the derived 'See' and 'Evoke' factor session data is held in the table "SessionD", and the "error" statistics and "skill" factors are held in the table "SessionK". All the data has been aggregated by user and image. In both datasets, the attribute "typeLanguage" is the “point of reference” for the data analysis and modeling. If the users' native language is not English, then 'typeLanguage'=2, whereas if the users native language is English, then 'typeLanguage' = 1. This indicator is used as the output, or labeling class.

5.1.1 Derived Factors (table SessionD)

Dataset 'SessionD' (Table 1) consists of a set of descriptive attributes about the tag session done by a user. The first two attributes have some information about the session they have just tagged: the image they found the easiest to tag and the image they found the most difficult to tag. The remaining factors represent the three concepts which we described previously in Section 3.4.2 of the paper; “Richness”, “Similarity” and “Spontaneity”. We recall that all these values are summarized by image and user, and that a tag consists of one or more terms (individual words).

"Richness" is represented by the four factors indicated by an ‘R’ in column 2 of Table 1. A description is given for each factor in column 3 of Table 1. The factors provide information related to the assignment of tags and terms, for ‘see’ and ‘evoke’ type tags.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SessionD: Derived dataset of semantic factors.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Factor</strong></td>
<td><strong>Type</strong></td>
</tr>
<tr>
<td>easiest</td>
<td>R</td>
</tr>
<tr>
<td>mostDifficult</td>
<td>R</td>
</tr>
<tr>
<td>NumTags_See</td>
<td>R</td>
</tr>
<tr>
<td>NumTags_Evoke</td>
<td>R</td>
</tr>
<tr>
<td>NumTerms_See1</td>
<td>R</td>
</tr>
<tr>
<td>NumTerms_Evoke1</td>
<td>R</td>
</tr>
<tr>
<td>Similarity_See1</td>
<td>SI</td>
</tr>
<tr>
<td>Similarity_Evoke1</td>
<td>SI</td>
</tr>
<tr>
<td>Similarity_See2</td>
<td>SI</td>
</tr>
<tr>
<td>Similarity_Evoke2</td>
<td>SI</td>
</tr>
<tr>
<td>Spontaneity_See</td>
<td>SP</td>
</tr>
<tr>
<td>Spontaneity_Evoke</td>
<td>SP</td>
</tr>
<tr>
<td>typeLanguage</td>
<td></td>
</tr>
</tbody>
</table>

*R=richness, SI=similarity, SP=spontaneity*
Table 2. SessionK: Derived dataset of syntactic factors

<table>
<thead>
<tr>
<th>Factors</th>
<th>Description*</th>
</tr>
</thead>
<tbody>
<tr>
<td>e0</td>
<td>Too many terms (more than four)</td>
</tr>
<tr>
<td>e1</td>
<td>Spelling mistakes</td>
</tr>
<tr>
<td>e2</td>
<td>Format wrong</td>
</tr>
<tr>
<td>e3</td>
<td>Wrong language</td>
</tr>
<tr>
<td>e4</td>
<td>False friends</td>
</tr>
<tr>
<td>etot</td>
<td>Total errors</td>
</tr>
<tr>
<td>IEC</td>
<td>Image evaluation coherence</td>
</tr>
<tr>
<td>Syntactic_level</td>
<td>Syntactic level factor based on errors</td>
</tr>
<tr>
<td>Semantic_level</td>
<td>Semantic level factor based on richness of vocabulary</td>
</tr>
<tr>
<td>Skill_level</td>
<td>Calculated from formula (1)</td>
</tr>
<tr>
<td>typeLanguage</td>
<td>Native or non-native</td>
</tr>
</tbody>
</table>

*also refer to the descriptions in Sections 4 and 5.1.2

"Similarity" is represented by the four factors indicated by an ‘SI’ in column 2 of Table 1. A description is given for each factor in column 3 of Table 1. The factors provide information related to the assignment of tags (averages and totals), for ‘see’ and ‘evoke’ type tags.

"Spontaneity" is represented by the two factors indicated by an ‘SP’ in column 2 of Table 1. A description is given for each factor in column 3 of Table 1.

5.1.2 Tagging Error Statistics: Errors made by users in tagging sessions

In dataset ‘SessionK’ (Table 2) we record five types of “error” made by the users during the tagging sessions. The error types have already been described in Section 4.

The first type of error, “description tag too long”, is designated as attribute ‘e0’ in table ‘SessionK’. The second type of error (orthographic) is designated as attribute ‘e1’. The third type of error (illegal characters) is designated as attribute ‘e2’. The fourth type of error occurs when the user writes the tag in a different language from the one defined for the tag session, for example the user writes “árbol” (Spanish) for “tree”. This is designated as attribute ‘e3’. This type of error was committed uniquely by the non native taggers. A fifth type of error, which is a subtype of error ‘e3’, occurs when the user incorrectly tries to adapt a word which s/he knows from his/her own language, thinking that it is similar to, or the same as the corresponding word in English. For example, the user writes “incomunication” (adaptation of a Spanish word) when the correct English word is “isolated”. This is designated as attribute ‘e4’. This type of error was committed uniquely by the non native taggers. Error types ‘e3’ and ‘e4’ are, of course, strong factors for discriminating between native and non-native taggers. The attribute ‘etot’ represents the total number of errors for a user, which is simply the sum of e0, e1, e2, e3 and e4.

Table 3. Percentage of the total number of users who commit the corresponding error type at least once, by type of language (Dataset ‘SessionK’)

<table>
<thead>
<tr>
<th>Error type</th>
<th>Native</th>
<th>Non native</th>
</tr>
</thead>
<tbody>
<tr>
<td>e0: too many terms</td>
<td>19%</td>
<td>4%</td>
</tr>
<tr>
<td>e1: spelling mistakes</td>
<td>73%</td>
<td>96%</td>
</tr>
<tr>
<td>e2: format wrong</td>
<td>38%</td>
<td>49%</td>
</tr>
<tr>
<td>e3: wrong language</td>
<td>0%</td>
<td>17%</td>
</tr>
<tr>
<td>e4: false friends</td>
<td>0%</td>
<td>13%</td>
</tr>
<tr>
<td>Etot: Total errors</td>
<td>78%</td>
<td>97%</td>
</tr>
</tbody>
</table>

With respect to the remaining attributes (Table 2, dataset ‘SessionK’), Syntactic_level is a derived factor based on the error frequencies, and IEC, Semantic_level are derived from the tagging session data in dataset ‘SessionD’, which have been described previously in Section 4. Finally, Skill_level is derived from Syntactic_level, IEC and Semantic_level, and has also been described in Section 4.

Table 3 summarizes the percentage of users who committed different type of errors. For example, we observe that 19% of all native users had one or more errors of type e0 (too many terms in the tag description). On the other hand, only 4% of the non native users displayed the same type of error. We observe that errors of type e3 and e4 were only committed by non native users. In total, 78% of all native users committed at least one error (of any type) and 97% of non native users. With respect to error types which discriminate between native and non native users, we would highlight e1, e3 and e4, although their frequencies of occurrence tend to be ‘sparse’.

6. Evaluation of descriptive statistics

Now we show results of the comparative data analysis based on the two derived datasets ‘SessionK’ and ‘SessionD’ described in Section 4 of the paper. For the task, we have used the IM4Data (IBM Intelligent Miner for Data V6.1.1) Data Mining tool [6]. The bivariate statistics function of IM4Data has been used for the data analysis, which generates summarized graphical representations, partitioned by the classifier category, which in the present study is ‘typeLanguage’ (native or non-native).

With reference to Figures 2, 3 and 4, we have used the ‘Bivariate statistics’ option of IM4Data, defining all variables as univariate with the exception of ‘typeLanguage’ which has been defined as bivariate (to calculate the Chi-Squared statistics relative to the other variables). In Figures 2, 3 and 4, the resulting graphics for each attribute are in descending order of the Chi-Squared statistic. Therefore, in Figure 3, attribute ‘e4’ has the highest Chi-Squared statistic in relation to the class label ‘typeLanguage’, followed by attribute ‘e3’, then ‘e1’, and so on. Numeric attributes are represented by a histogram of their distribution, and categorical attributes are represented by a ‘pie’ chart. In all the histograms, the grey filled rectangles indicate the frequency for the whole dataset, and the other rectangles (which terminate above or below the grey rectangles, represent the frequency for the given partition.
6.1 Evaluation of Semantic Factors

Figure 2 has been produced from the ‘SessionD’ dataset for native English taggers (left column of variables) and non-native taggers (right column of variables). This dataset contains attributes which represent the ‘richness’, ‘similarity’ and ‘spontaneity’ factors for the user tag sessions. In Figure 2, we show the first four variables ranked by the Chi Squared statistic, which relates them to the type of user (native or non native). Two of the attributes most related to the native/non native label (as indicated by Chi-Squared) are variables related to the similarity of the evoke type tags: ‘Similarity_Evoke2’ and ‘Similarity_Evoke1’. This agrees with the hypothesis that non native users have more difficulty to think of vocabulary which defines emotions.

![Figure 2. Distributions of selected semantic factors, for native (left) and non native (right) taggers.](image)

The distributions of ‘Similarity_Evoke2’ and ‘Similarity_Evoke1’ in Figure 2 show that the non-natives (histogram on the right) have a greater frequency in the higher (rightmost) part of the distribution, which means that there is more coincidence between the non-native tags, and therefore less diversity. The ‘mostDifficult image’ attribute shows a significant difference between which images the native and non native taggers found the most difficult. Another significant tendency is found for attribute ‘NumTerms_Evoke1’. Refer to the descriptions of the semantic factors in Table 1 of Section 5.1.1. If we compare the distributions on the right/left for this attribute, we see on the right (non native) a significantly greater frequency for more terms per tag. This means that non natives are using more terms for evoke type tags, to express a concept. On the other hand, natives are finding more exact terms to describe the images.

6.2 Evaluation of Syntactic Factors

Figures 3 and 4 have been produced from the ‘SessionK’ dataset for native English taggers (left column of attributes) and non-native taggers (right column of attributes). We recall that this dataset contains attributes which represent the error rates for different types of errors, and the derived ‘skill level’ attributes and factors. We can visually compare the distributions of the different variables, for native (left column) and non-native (right column) users, identifying any trends. Attributes ‘e4’ and ‘e3’ appear first in the Chi-square ordered attributes, and are therefore the most discriminatory between native and non-native. This is because ‘e4’ represents the ‘false friend’ error count and ‘e3’ represents the wrong language error count, and these errors were only made by non-natives. We note that the error frequencies are ‘sparse’, that is, many users have no errors of a given type (although a majority of users have at least one error of any type, see Table 3). For attribute ‘e4’ for native users (Figure 3, left column) 100% of the users (left bar of the histogram) had zero errors of type e4, and 0% of the users had one error of type e4 (right bar of the histogram). The grey area indicates that the overall percentage of users with zero errors of type e4 is 85%, and the overall percentage of users with one error of type e4 is 15%. If we look at the distribution of ‘e4’ for non native users (Figure 3, right column), we observe that 70% (left bar of the histogram) had zero errors of this type, and 30% had one error of this type. The grey filled rectangles represent the overall frequencies for whole dataset.

In the case of attribute ‘e1’ of Figure 3 (natives, left histogram), if we compare with the same attribute for non natives (right histogram), we observe a significantly greater frequency for the leftmost bar of the histogram. This indicates that a greater number of native taggers make a smaller number of spelling mistakes.

For attribute ‘etot’, the total number of errors of all types committed, we see a general tendency where native taggers have higher incidence for 0 and 1 errors (left histogram), whereas non-native taggers have a higher incidence for 2, 3 and 4 errors. We also observe a small exceptional group of native taggers with 7 or more errors (left histogram).

6.3 Evaluation of Skill Factors

With reference to Figure 4, we see the histogram distribution for attribute ‘skill level’ and the three skill factors which are used to calculate the skill level. With reference to the first attribute,
‘skill_level’, in the histogram on the left (native taggers), we observe that there are higher frequencies in the rightmost part of the distribution (skill value > 4.5), when compared to the distribution of ‘skill_level’ in the histogram on the right (non-native taggers). This confirms that a slightly higher general skill level is shown by the natives with respect to the non-natives. Also, in the distribution of ‘skill_level’ in Figure 4 (right), we observe higher frequencies in the leftmost part of the distribution (skill value < 3.5), when compared to the distribution of ‘skill_level’ in Figure 4 (left).

With respect to the other skill factors, ordered by Chi-Square statistic, we observe that native taggers (left histogram) have a lower syntactic level (higher frequencies in the low end of the range), which, following our definition, means they commit less errors. In the case of the semantic level, the native taggers have higher frequencies in the high/ end of the range (which for this factor means that they have defined more distinct tags and therefore have a richer vocabulary). Non native users, on the other hand, have greater frequencies in the middle/lower range of semantic level, which indicates that they tend to use less tags. Finally, we observe that attribute, IEC does not discriminate between native and non-native, although we can observe that natives have a smaller percentage of low IEC values (65% with respect to 70%) which means that they make slightly less mistakes in correctly identifying the easiest and most difficult images to tag.

7. TAG RECOMMENDATION BASED ON USER TYPE AND ERROR TYPE

Now we can summarize the different responses and recommendations we would give for each of the user and error types. For error types 1, 2 and 3, the user can be given a simple message indicating the error and with possible automatic correction (see Section 5.1.2).

Error type 4: tag in incorrect language. The user writes “arbol” (spanish) for “tree”. The system does not find the word in the English dictionary. One option would be to have established the native language of the tagger during the calibration session. Then the system would look in the appropriate dictionary for the translation to English.

Error type 5: incorrect adaptation of a foreign word. The user writes “exit” as the tag for an image, thinking it means ‘success’ because the Spanish word for ‘success’ is ‘éxito’. The procedure to detect the error cannot be the same as for error type 4, because the word is a valid English term and therefore the system will find
it in the English dictionary. One option would be to have an extensive list of words which are susceptible to this type of error, by taggers in English whose native language is Spanish. Then the system would detect the word “exit” as one of these.

Recommendation of ‘evoke’ tags based on ‘see’ tags: if the user has already defined the ‘see’ tags, then the system can recommend the ‘evoke’ tags, based on the ‘see’ tags. For example, with reference to the list of most frequent ‘see’ and ‘evoke’ tags for Image 10, which we saw in Table 4, the non-native user defines the following ‘see’ tags: ‘sky’, ‘grass’ and ‘bench’. Then the system would consult a dictionary of ‘see’ tags and corresponding ‘evoke’ tags which have been defined previously by other (native or more highly skilled) users. In the case of ‘sky’, we see that the system would retrieve ‘nature’, which is the native users corresponding tag to ‘sky’. Similarly, the system would retrieve ‘peace’ (evoke) for ‘bench’ (see) and ‘open’ (evoke) for ‘grass’. It can be debated if the resulting tags are linguistically better, more expressive or richer than the evoke tags the user may have defined him or herself. Also, it may happen that the ‘see’ tag the user has defined does not exist in the system’s dictionary.

User support based on derived factor ‘skill_level’ : with reference to Figure 2, and the distribution of the attribute ‘skill_level’, we observe the leftmost part of the distribution (< 3.5) has a low skill level. We can offer specific additional support to these users. The middle range of the distribution (between 3.5 and 4.5) represents the users with average skill level, who could have ad-hoc support and advice depending on the type of errors made or ‘see’/‘evoke’ suggestions. Lastly, the rightmost part (> 4.5) of users with the highest calculated skill level can be used as a ‘model’ for other users, for example, by saving their ‘see’ and ‘evoke’ tags for future reference.

8. CONCLUSIONS AND FUTURE WORK
In this study we have analyzed two types of factors which influence in the tagging process: (i) semantic factors such as “richness”, “similarity” and “spontaneity”, which we have captured for “objective” tagging and “subjective” tagging; (ii) syntactic factors defined as types of errors that users commit during tagging. We have related these factors for native and non-native taggers in the English language.

Based on the analysis of the descriptive statistics in Section 6, we can reasonably conclude that: (a) Natives have better tagging skills than non-natives; (b) Subjective tagging is more difficult than objective tagging and therefore requires more support for the user. Therefore we have confirmed our two initial hypothesis.

Thus, native and non-native taggers have distinctive characteristics in terms of the tag type (especially ‘evoke’) and in terms of some, but not all of the error types. In terms of error types, there is an overlap for natives and non-natives, with the exception of ‘wrong language’ and ‘false friend’ type errors, although the non-natives generally tend to have higher error rates (Table 3).

As a next step in our work, we propose testing natives of other languages (such as German and Mandarin Chinese) as non-native taggers in English. These languages, Teutonic and Oriental based, respectively, would be an interesting contrast to the linguistic peculiarities of the Spanish (Latin based) non-native taggers.

9. REFERENCES

Multilingual Wikipedia, Summarization, and Information Trustworthiness

Elena Filatova
Fordham University
Department of Computer and Information Sciences
filatova@cis.fordham.edu

ABSTRACT

Wikipedia is used as a corpus for a variety of text processing applications. It is especially popular for information selection tasks, such as summarization feature identification, answer generation/verification, etc. Many Wikipedia entries (about people, events, locations, etc.) have descriptions in several languages. Often Wikipedia entry descriptions created in different languages exhibit differences in length and content. In this paper we show that the pattern of information overlap across the descriptions written in different languages for the same Wikipedia entry fits well the pyramid summary framework, i.e., some information facts are covered in the Wikipedia entry descriptions in many languages, while others are covered in a handful number of descriptions. This phenomenon leads to a natural summarization algorithm which we present in this paper. According to our evaluation, the generated summaries have a high level of user satisfaction. Moreover, the discovered pyramid structure of Wikipedia entry descriptions can be used for Wikipedia information trustworthiness verification.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing

General Terms

Measurement, Experimentation, Human Factors

Keywords

Wikipedia, summarization, multilinguality

1. INTRODUCTION

Wikipedia\(^1\)\(^2\) provides descriptions of people, events, locations, etc. in many languages. Despite the recent discussion of the Wikipedia descriptions trustworthiness or lack of thereof [9], Wikipedia is widely used in information retrieval (IR) and natural language processing (NLP). The question arises: what can be done to increase the trustworthiness of the information extracted from Wikipedia. We believe, Wikipedia itself has resources to increase its trustworthiness.

Most Wikipedia entries have descriptions in different languages. These descriptions are not translations of a Wikipedia entry description from one language into other languages. Rather, Wikipedia entry descriptions in different languages are independently created by different users. The length of the entry descriptions about the same Wikipedia entry varies greatly from language to language. Obviously, texts of different length cannot contain the same amount of information about an entry.

In this paper we compare descriptions of Wikipedia entries written in different languages and investigate the information overlap pattern in multilingual Wikipedia. We show that information overlap in entry descriptions written in different languages corresponds well to the pyramid summarization model [16, 12]. This result helps the understanding of the combined value of the multilingual Wikipedia entry descriptions. On the one hand, multilingual Wikipedia provides a natural summarization mechanism. On the other hand, to get a complete picture about a Wikipedia entry, descriptions in all languages should be combined. Finally, this pyramid structure can be used for information trustworthiness verification.

The rest of the paper is structured as follows. In Section 2 we describe related work. In Section 3 we provide a motivation example for our research. In Section 4 we describe our corpus, the summarization-based experiments we ran to analyze multilingual Wikipedia information overlap. In Section 5 we discuss the results of our experiments. In Section 6 we outline the avenues for future research.

2. RELATED WORK

Multilingual aspect of Wikipedia is used for a variety of text processing tasks. Adafre et al. [8] analyze the possibility of constructing an English-Dutch parallel corpus by suggesting two ways of looking for similar sentences in Wikipedia pages (using matching translations and hyperlinks). Richman et al. [13] utilize multilingual characteristics of Wikipedia to annotate a large corpus of text with Named Entity tags. Multilingual Wikipedia is used to facilitate cross-language IR [14] and to perform cross-lingual QA [6].

The described applications do not question the trustworthiness of the information presented in Wikipedia. In a separate line of research, approaches are developed to rate the trustworthiness of Wikipedia information. These approaches, however, deal with the text only one language.

Wikipedia content trustworthiness can be estimated using a combination of the amount of the content revision and the author reputation performing this revision [2]. Another way to use edit history to estimate information trustworthiness is to use dynamic Bayesian network trust model that utilized rich revision information in Wikipedia. [17]. Wikipedia trustworthiness can be also estimated using an additional tab (Trust tab) [11].

---


\(^2\) Wikipedia is changing constantly. All the quotes and examples from Wikipedia presented and analyzed in this paper were collected on February 10, 2009, between 14:00 and 21:00 PST.
The research closest to ours was recently described in Adar et al. [1] where the main goal is to use self-supervised learning to align or/and create new Wikipedia infoboxes across four languages (English, Spanish, French, German). Wikipedia infoboxes contain a small number of facts about Wikipedia entries in a semi-structured format. In our work, we deal with plain text and disregard any structured data such as infoboxes, tables, etc. It must be noted, that the conclusions that are reached in parallel for structured Wikipedia information by Adar et al. and for unstructured Wikipedia information by us are very similar. These conclusions stress the fact that the most trusted information is repeated in the Wikipedia entry descriptions in different languages. At the same time, no single entry descriptions can be considered as the complete source of information about a Wikipedia entry.

3. INFORMATION OVERLAP

Currently, Wikipedia has entry descriptions in more than 200 languages. The language with the largest number of entry descriptions is English [8, 5] but the size of non-English Wikipedia is growing fast and represents a rich corpus. Most existing NLP applications that use Wikipedia as the training corpus or information source assume that Wikipedia entry descriptions in all languages are a reliable source of information. However, according to our observations, Wikipedia descriptions about the same entry (person, location, event, etc.) in different languages frequently cover different sets of facts. According to the Wikipedia analysis [7], there are two major sources of differences in the descriptions of the same Wikipedia entry written in different languages:

- the amount of information covered by a Wikipedia entry description;
- the choice of information covered by a Wikipedia entry description.

In this paper we analyze the information overlap in Wikipedia entry descriptions written in several languages.

For example, baseball is popular in the USA, Latin America, and Japan but not in Europe or Africa. Wikipedia has descriptions of Babe Ruth in 18 languages: the longest and most detailed descriptions are in English, Spanish and Japanese. The description of Babe Ruth in Finnish has five and in Swedish - four sentences. These short entry descriptions list several general biographical facts: dates of birth, death; the fact that he was a baseball player. It is likely, that the facts from the Swedish and Finnish entry descriptions about Babe Ruth will be listed in a summary of the English language Wikipedia entry description of him.

Currently, information overlap is successfully used for summarization evaluation. The state-of-the-art (automatic and manual) summarization evaluation approaches compare the target summary against several model summaries. Such models are typically created manually. The more model summaries contain a specific piece of information - the greater value it gets in the target summary [10, 16, 12].

4. CORPUS ANALYSIS EXPERIMENT

In this paper, we investigate how the information overlap in multilingual Wikipedia can be used to create summaries of entry descriptions.

4.1 Data Set

For our experiment, we used the list of people created for the Task 5 of DUC 2004: biography generation task (48 people). This set is small enough to be analyzed manually in detail; at the same time, it is widely used for summarization experiments. In the future, we plan to compare the performance of our summarization approach against another summarization system developed for biography generation that utilizes Wikipedia document structure [4].

We downloaded from Wikipedia all the entry descriptions in all the languages corresponding to each person from the DUC 2004 list. We used Wikitext, the text that is used by Wikipedia authors and editors. Wikitext can be obtained through Wikipedia dumps. We removed from the wikitext all the markup tags and tabular information (e.g., infoboxes and tables) and kept only plain text. There is no commonly accepted standard wikitext language, thus our final text had a certain amount of noise which, however, as discussed in Section 5, did not affect our experimental results.

For each Wikipedia entry (i.e., DUC 2004 person) we downloaded corresponding entry descriptions in all the languages, including Esperanto, Latin, etc. We used the name of a person to find the description of this person in English and then we followed the links from the left side panel of the Wikipedia page template to get the entry descriptions in other languages. To facilitate the comparison of entry descriptions written in different languages we used the Google machine translation tool to translate the downloaded entry descriptions into English. The number of languages covered currently by the Google translation system (41) is less than the number of languages used in Wikipedia (265). However, the language distribution in the collected corpus corresponds well the language distribution in Wikipedia and the collected Wikipedia subset can be considered a representative sample [7].

Five people from the DUC 2004 set had only English Wikipedia entry descriptions: Paul Coverdell, Susan McDougal, Henry Lyons, Jerri Nielsen, Willie Brown. Thus, they were excluded from the analysis. The person whose Wikipedia entry had descriptions in most languages (86) was Kofi Annan. On average, a Wikipedia entry for a DUC 2004 person had descriptions in 25.35 languages. The description in English was not always the longest description: in 17 cases the longest description of a Wikipedia entry for a DUC 2004 person was in a language other than English.

4.2 Data Processing Tools

After the Wikipedia entry descriptions for the DUC 2004 set were collected and translated, we divided these descriptions into sentences using the LingPipe sentence chunker [3]. For each DUC 2004 person we compared a description of this person in English against the descriptions of this person in other languages that were handled by the Google translation system. We counted descriptions in how many languages had sentences corresponding to the sentences in the description in English. To identify matching sentences we used the LingPipe string matching tool based on TF-IDF distance which “is based on vector similarity (using the cosine measure of angular similarity) over dampened and discriminatively weighted term frequencies. [...] two strings are

3 http://meta.wikimedia.org/wiki/List_of_Wikuseridias
4 In this work, the length of a Wikipedia entry description is measured in sentences used in the text description of a Wikipedia entry.
6 http://download.wikimedia.org/
7 http://translate.google.com/
### 4.3 What was Measured

To evaluate how much information is repeated in the descriptions of the same person in different languages, we measured the similarity of the person's description in English and in other languages. As each sentence was treated as a separate document, the number of tokens (words) for comparison was rather small. Thus, for the 0.5 similarity threshold, the sentences marked as similar were almost identical. Using the 0.35 and 0.2 thresholds allowed to search for non-identical sentences that still had a substantial word overlap.

Our hypothesis is that those facts (sentences) that are mentioned in the descriptions of a person in different languages fit well the pyramid summarization model. For example, if we are to summarize a description of a person from the English Wikipedia: first, we should add to the summary those sentences that have their counterparts in the most number of descriptions of this person in the languages other than English. Sentences added on this step correspond to the top level of the pyramid. If the length of the summary is not exhausted then, on the next step, we add to the summary those sentences that appear in the next most number of languages, and so on. Thus, we can place sentences on different levels of the pyramid, with the top level being populated by the sentences that appear in the most languages and the bottom level having sentences that appear in the least number of languages. For our experiments we used sentences from the top three levels of this pyramid. All the sentences added to the summary should appear in at least two languages other than English. Table 1 has a schematic outline of the described algorithm.

<table>
<thead>
<tr>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Submit the person's name to Wikipedia</td>
</tr>
<tr>
<td>2. Get Wikipedia entry descriptions for this person in all possible languages</td>
</tr>
<tr>
<td>3. Remove plain text information from the descriptions</td>
</tr>
<tr>
<td>4. For all the languages handled by the Google MT, translate entry descriptions into English</td>
</tr>
<tr>
<td>5. Break English texts into sentences</td>
</tr>
<tr>
<td>6. Use a similarity measure to identify what English sentences have counterparts in entry descriptions in other languages</td>
</tr>
<tr>
<td>7. Rank all the sentence from the English document according to the number of languages that have similar sentences</td>
</tr>
<tr>
<td>8. If several sentences are placed on the same level, list these sentences in the order they appear in the Wikipedia entry description in English</td>
</tr>
<tr>
<td>9. Use the top three levels from the above ranking</td>
</tr>
</tbody>
</table>

Table 1: Algorithm outline.

---

...more similar if they contain many of the same tokens with the same relative number of occurrences of each. Tokens are weighted more heavily if they occur in few documents" [3]. For our experiment, each sentence was treated as a separate document. The IDF value was computed based on the two entity descriptions under consideration (one - in English, the other one - translation into English). We used three similarity thresholds: 0.5, 0.35, 0.2.

The summary created using the 0.2 threshold contains the introductory sentence as well as sentences not included in the summaries for the 0.5 and 0.35 similarity threshold.

Despite the fact that for our experiment we chose the set of people used for the DUC 2004 biography generation task, we could not use the DUC 2004 model summaries for our evaluation. These models were created using the DUC 2004 corpus, while in our experiments we used a subset of multilingual Wikipedia. Moreover, Wikipedia entry descriptions about the DUC 2004 people had dramatic updates since 2004. For example, Jörg Haider died of injuries from a car crash on October 11, 2008 and this information is included into our three-level summaries. Due to space constraints, in this paper we report only the results obtained using similarity threshold of 0.35. Also, in the experiment described in this paper we analyze only those sentences from the English text that appear in at least two other languages, with the exception for Louis Frech, for whom only one language was handled by the Google Translation system. Thus, the summary for the English entry description about Louis Frech has only one level which has all the sentences from the English entry description that have their counterparts in the only available translation. Thus, using the DUC 2004 set we created:

- one-level summaries for 5 people;
- two-level summaries for 3 people;
- three-level summaries for 35 people.

The length of the created summaries is measured in sentences. Table 3 presents information about the average and maximal length of summaries for all three levels combined and for each level separately. The summaries that do not have Level 2 and/or Level 3 are excluded from the corresponding average and maximum value computation. According to the presented data, on average, the output three-level summaries are rather short, however, some summaries can be quite long. We believe that such a difference between the average and the maximal length is due to:

1. the length variation of the English Wikipedia entry descriptions;
2. the number variation of descriptions (languages) for each a person and the lengths of these descriptions.

To evaluate the output three-level summaries we used Amazon Mechanical Turk as a source of human subjects who can reliably evaluate certain NLP tasks [15]. For each of the 43 outputs (for 43 people from the DUC 2004 set) we
Lang. | Sent. ID | Text
---|---|---
1 | 3 | Orvon Gene Autry (September 29, 1907 – October 2, 1998) was an American performer, who gained fame as The Singing Cowboy on the radio, in movies and on television.

Similarity 0.5

1 | 7 | Orvon Gene Autry (September 29, 1907 – October 2, 1998) was an American performer, who gained fame as “The Singing Cowboy” on the radio, in movies and on television.
2 | 3 | Autry, the grandson of a Methodist preacher, was born near Tioga, Texas. His first hit was in 1932 with “That Silver-Haired Daddy of Mine,” a duet with fellow railroad man, Jimmy Long.
3 | 2 | After leaving high school in 1925, Autry worked as a telegrapher for the St. Louis-San Francisco Railway.
4 | 13 | Autry also sang the classic Ray Whitley hit “Back in the Saddle Again,” as well as many Christmas songs including “Santa Claus Is Coming to Town,” his own composition “Here Comes Santa Claus,” “Frosty the Snowman,” and arguably his biggest hit “Rudolph the Red-Nosed Reindeer.”
5 | 72 | Gene Autry died of lymphoma at age 91 at his home in Studio City, California and is interred in the Forest Lawn, Hollywood Hills Cemetery in Los Angeles, California.

Similarity 0.35

1 | 7 | Orvon Gene Autry (September 29, 1907 – October 2, 1998) was an American performer, who gained fame as “The Singing Cowboy” on the radio, in movies and on television.
2 | 6 | Autry, the grandson of a Methodist preacher, was born near Tioga, Texas. His death on October 2, 1998 came nearly three months after the death of another celebrated cowboy of the silver screen, radio, and TV, Roy Rogers.
3 | 5 | From 1940 to 1956, Autry had a huge hit with a weekly radio show on CBS, “Gene Autry’s Melody Ranch.” His horse, Champion, also had a radio-TV series “The Adventures of Champion.” Gene Autry died of lymphoma at age 91 at his home in Studio City, California and is interred in the Forest Lawn, Hollywood Hills Cemetery in Los Angeles, California.

Similarity 0.2

1 | 7 | Orvon Gene Autry (September 29, 1907 – October 2, 1998) was an American performer, who gained fame as “The Singing Cowboy” on the radio, in movies and on television.
2 | 6 | Autry, the grandson of a Methodist preacher, was born near Tioga, Texas. His death on October 2, 1998 came nearly three months after the death of another celebrated cowboy of the silver screen, radio, and TV, Roy Rogers.
3 | 5 | From 1940 to 1956, Autry had a huge hit with a weekly radio show on CBS, “Gene Autry’s Melody Ranch.” His horse, Champion, also had a radio-TV series “The Adventures of Champion.” Gene Autry died of lymphoma at age 91 at his home in Studio City, California and is interred in the Forest Lawn, Hollywood Hills Cemetery in Los Angeles, California.

Table 2: Three-level summaries for Gene Autry (#: the summary level; Lang.: number of languages that contain a similar sentence; Sent. ID: the position of the sentence in the English description; Text: the sentence itself).

<table>
<thead>
<tr>
<th>Three-level summary</th>
<th>Level one</th>
<th>Level two</th>
<th>Level three</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg</td>
<td>3.74</td>
<td>1.92</td>
<td>1.58</td>
</tr>
<tr>
<td>Max</td>
<td>9</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 3: Summaries length: average and maximal.

We recruited five human annotators. The annotators were provided with: the name of the person; link to the Wikipedia entry description about this person in English; three-level summary of this Wikipedia entry description. We asked our human annotators to answer the following questions:

- Do you agree that the sentences listed on Level 1 are a good summary of the Wikipedia entry description about Person (assume, the number of sentences in the summary cannot exceed the number of sentences listed on Level 1)?
- Assume that the summary of the Wikipedia entry description about Person can have as many sentences as listed on Level 1 and Level 2 combined. Do you agree that the sentences listed on Level 1 and Level 2 are a good summary?
- Assume that the summary of the Wikipedia entry description about Person can have as many sentences as listed on Level 1, Level 2, and Level 3 combined. Do you agree that the sentences listed on Level 1, Level 2, and Level 3 are a good summary?

If the summary did not have Level 2 and/or Level 3 sentences, the annotator was asked to skip answering the corresponding questions.

5. RESULTS

Table 4 summarizes the results of the three-level summaries evaluation. The Goodness measure shows how many (out of five) annotators agreed that the summary for a particular level, given the length constraint, was good. The numbers in the table show the number of summaries that were considered good for each level according to a particular level of goodness. As it is mentioned in Section 4.4 not all summaries have Levels 2 and 3 filled in; the Number of summaries column in Table 4 has this information.

According to Table 4, no summary on Level 1 was uniformly considered bad. One summary was considered bad by four out of five annotators. This was the summary for Paul Wellstone with Level 1 consisting only of one sentence. We analyzed this sentence and discovered that it was incorrectly truncated due to our sentence chunker error.

[Paul David Wellstone (July 21, 1944 - October 25, 2002) was a two-term U.S. Senator from the U.S. state of Minnesota and member of the Democratic-Farmer-Labor Party, which is affiliated with the national Democratic Party.]

This sentence was broken into two sentences (identified above by the square brackets), and only the first portion of the sentence was added to the Level 1 summary. Despite the fact that this portion contains important biographical information, it cannot be used as a stand-alone sentence. According to our analysis, three out of seven summaries that were judged as bad by three out of five annotators had exactly the same problem of incorrect sentence segmentation forcing only portions of sentences to be added to the summaries.

In addition to asking annotators to judge the quality of the created summaries we welcomed our annotators to leave comments about the summaries they read. These comments can be divided into two groups. Several annotators noticed text preprocessing errors (e.g., leftovers from the Wikitext XML tagging); however, this did not affect their judgement of the summary quality: all the summaries containing XML tags were marked as good. The other set of observations...
concerned the type of facts included in the summaries. For example, it was pointed out that the sentences from the summary about Abdullah Ocalan did not have enough information about his political activities and thus, the created summaries were judged as bad. Annotators suggested that information about professional life of politicians would be more appropriate than the information about their marriages. However, sentences containing information about private life were mostly considered relevant and judged as good additions to summaries.

Figure 1 shows the combined numbers for Table 4. For each level we grouped all the numbers in two categories: those numbers where the majority of the annotators agreed that the summary was good and those numbers where the majority of the annotators decided that the summary was bad. As not all the summaries had sentences from all three levels, Figure 1 has ratios rather than the absolute numbers listed in Table 4. This figure shows that overall the quality of the created summaries was quite high. In more than 80% of cases our annotators were happy with the summaries consisting of the Level 1 sentences, and in more that 70% of cases our annotators were happy with the summaries consisting of the sentences combined from Levels 1 and 2 and Level 1, 2, and 3. To conclude this section, we showed that information overlap in multilingual Wikipedia can be used for placing information facts into a pyramid structure.

### 6. FUTURE WORK

While the main focus of the current paper is information overlap, in the future, we are interested in studying information asymmetries in multilingual Wikipedia.

We are interested in investigating how Wikipedia multilinguality can be used for opinion, contradiction and new information detection. An important observation concerning the example presented in Section 3 is that irrespectively of the length of the descriptions of a person in different languages, none of these descriptions have any facts that contradict the facts in the descriptions of this Wikipedia entry in other languages. Rather, the discussed entry descriptions in different languages contain a subset of facts that appear in many languages plus, maybe, additional information. This allowed us to formulate and test the hypothesis that a set of Wikipedia entry descriptions about the same entry fits well the pyramid summarization model. However, there are Wikipedia entries attitude to which is different in communities speaking in different languages. In these cases, we believe, it will be more interesting to draw the readers attention not to the facts that are repeated in the entry descriptions in many languages but rather, highlight differences among these descriptions.

### 7. REFERENCES


NSContrast: An Exploratory News Article Analysis System that Characterizes the Differences between News Sites

Masaharu Yoshioka
Graduate School of Information Science and Technology, Hokkaido University
National Institute of Informatics
N14 W9, Kita-ku, Sapporo-shi
Hokkaido Japan
yoshioka@ist.hokudai.ac.jp

ABSTRACT
The News Site Contrast (NSContrast) system analyzes multiple news sites using the concept of contrast set mining and can extract the terms that characterize the differences in topics of interest for each country. However, because of the poor quality of some machine translation, NSContrast results include some meaningless terms generated by this mistranslation. To address this problem, Wikipedia is used as a bilingual dictionary and as a source for synonym identification. We give some experimental results for this New NSContrast system.

Categories and Subject Descriptors
H3.3 [Information Systems]: Information Search and Retrieval

General Terms
Information Retrieval, Text Mining

Keywords
News search, IR interface, Contrast set mining

1. INTRODUCTION
It has recently become possible to access a wide variety of news sites from across the world via the Internet. Because each news site has its own culture and interpretation of events, we can obtain a greater diversity of information using multiple news sites than ever before.

Because each country has different opinions and interests, when we use news sites from different countries, we will obtain different points of view for a topic. For example, considering diplomatic issues to do with North Korea; Asian, European and American news sites have some common interests and their own characteristic interests. Therefore, to analyze events reported from multiple sites, it is important to clarify the characteristics of each news site.

There are several experimental systems that integrate news articles about a particular event from multiple news sites. For example, EMM News Explorer 1 and Newsblaster 5 are integrated news aggregation systems from distributed news archives.

These systems are effective for understanding a particular event via multiple news sites, but they ignore the characteristics of each news site. For example, Japanese news sites tend to report Japanese-related topics more frequently than others. To better understand articles from different news sites, this bias should be taken into account. To identify news site characteristics, NSContrast 10 has been proposed. This system analyzes multiple news sites using the concept of contrast set mining and aims to extract the characteristic information about each news site by performing term co-occurrence analysis. The system has potential for extracting characteristic terms that reflect topic divergence between different countries. However, because of the poor quality of some machine translation, NSContrast results include some meaningless terms generated by this mistranslation 9.

In this paper, we discuss the issues related to handling articles from multiple news sites via machine translation, and we propose a method for using Wikipedia as a resource for constructing a bilingual dictionary for an NSContrast system. Using this dictionary, a new database of news articles is constructed, which is then tested via user experimentation with an NSContrast system.

2. NSCONTRAST

2.1 Term Collocation Analysis by Contrast Set Mining
Term collocation analysis is a well-known text mining method for extracting characteristic information from texts[7]. However, conventional collocation analysis mostly focuses on the characteristic information that is dominant in the text database. In many cases, most of the information is well known and is therefore not particularly interesting.

To solve this problem, we introduce the concept of contrast set mining 8 for the analysis. This framework compares a global data set and a conditioned local data set to find characteristic item information that is significantly different from the global characteristic information. Even though this information may not be dominant in either the global or the local data set, it can be used to understand the

1http://press.jrc.it/NewsExplorer/
characteristics of the local database.

We use Discovery of Correlation (DC) pair mining [8] for term collocation analysis. In DC pair mining, the “difference in correlations observed by conditioning a local database” is of particular interest. To quantify this difference, we introduce a new measure, \( change(X, Y; C) \), defined by

\[
change(X, Y; C) = \frac{correl(X, Y)}{correl(X, Y_i)}
\]

where \( X \) and \( Y \) represent the item sets and \( C \) represents the condition for creating the local database. \( correl(X, Y) \) and \( correl(X, Y_i) \) correspond to the correlations between \( X \) and \( Y \) in the global database and a \( C \)-conditioned local database, respectively.

To utilize this technique for term collocation analysis of multiple text databases (news text databases from different countries), we modify this analysis method as follows.

- Because the size of each text database is not much smaller than that of the global database, a characteristic collocation occurring only in one database is also a characteristic collocation in the global database. Therefore, the contrast between the conditioned database and the rest of the databases is used instead of the original definition.
- Because the computational cost of DC pair mining is substantial, the target term for analysis (\( X \) in the formula) is given by the user.

By using this technique, we can extract characteristic minor topics that are of interest (higher change : \( X_n \) in Figure 1) or are neglected (lower change : \( X_m \) in Figure 1) in one database compared with others.

![Collocation Analysis based on DCPair Mining](image)

**2.2 NSContrast: A News Site Analysis System**

The NSContrast system is a method for accessing news articles from multiple news sites using the concept of DC pair mining[10, 9]. This system has the following analytic components.

- Term collocation analysis based on DC pair mining.
  - The system generates a list of characteristic terms by comparing news article databases from different countries. This term list is represented as a term collocation graph to aid understanding of the relationships among characteristic terms.
  - A burst analysis function [4] for finding an appropriate time sequence window.
  - To find good characteristic terms using contrast set mining, it is preferable to select a large number of articles for a particular topic. Because burst analysis is a method that finds a period during which a given term is of more interest than usual, it is an effective technique for finding this information.
  - A news article retrieval system.
  - To understand the meaning of term collocation analysis and burst analysis, a news article retrieval system is used for this purpose.

3. A NEW NSCONTRAST USING WIKIPEDIA INFORMATION

We constructed a news article database by collecting news articles from news sites in Japan, China, Korea, and the USA. We confirmed that NSContrast has capabilities that extract characteristic terms for understanding the differences between news articles on a particular topic from multiple countries.

However, during experiments, a user claimed that the system tends to select the following two types of meaningless characteristic terms.

1. Terms generated by mistranslation.
   - For example, “奥巴马” means Obama (President of the United States) in Chinese, but the machine translation system translates this term as “オーストラリア” (President of Australia). Because of this mistranslation, NSContrast tends to select “オーストラリア (Australia)” as a characteristic term for China when the system is analyzing articles for a particular topic.

2. Terms with many synonyms.
   - For example, “北朝鲜 (North Korea)” , “北朝鮮民主主義人民共和国 (Democratic People’s Republic of Korea)” , and “DPRK” are equivalent terms for the same country in Asia. When one of the news sites uses a different representation from others, this term tends to be selected as a characteristic term.

From an analysis of these problems, we found that they occur mainly when handling named entities (e.g., names of people, countries, and companies) in a Chinese–Japanese translation system. This is because the dictionary for the system is updated only infrequently and therefore lacks entries for recently named entities.

Because Wikipedia has many entries related to named entities and these entries are associated via language links (equivalent relationships among different languages) and redirection links (reference relationships in the same language), it is a good resource for constructing both bilingual dictionaries and synonym dictionaries.

In this research, we constructed a Chinese–Japanese bilingual dictionary from Wikipedia, based on the method proposed by [2]. Because it is easier for the machine translation system to translate Japanese named entities that are described in terms of Chinese characters than katakana’s one
(katakan contains phonograms that are used mainly to represent terms imported from other countries), we constructed a bilingual dictionary for Japanese katakana terms, as follows.

From a Japanese Wikipedia data dump, we selected Japanese katakana entries that have a language link to Chinese entries.
For example, “バラク・オバマ (Barack Obama)” is selected because it is a katakana entry and has a Chinese language link to “巴郎克·奧巴馬”.

2. Separation of first name and family name.
For most people, in both katakan and Chinese, the first name and the family name are separated by “・”.
The correspondence between these names is specified by using “・”.
For example, from a language link “バラク・オバマ (Barack Obama)” to “巴郎克·奧巴馬”, two dictionary entries (“巴郎克” to “バラク (Barack)” and “奧巴馬” to “オバマ (Obama)”) are generated.

In the Chinese language, two or more different Chinese expressions for the same person often exist. In such cases, redirection links in the Chinese Wikipedia are used to represent the relationship among these entries.
By using these links, bilingual dictionary entries for handling these terms were generated.
For example, from a Chinese redirection link “歐巴馬” to “奧巴馬” and a language link “奧巴馬” to “オバマ (Obama)”, a new dictionary entry “歐巴馬” to “オバマ (Obama)” is generated.

By using this method, 13,255 bilingual dictionary entries (Chinese to Japanese) were extracted from Japanese and Chinese Wikipedia data dumps.

We used this Chinese-Japanese dictionary as an add-on external bilingual dictionary for a machine translation system by using the Language grid service [1, 6] that supports the combination of a dictionary lookup service with a machine translation system.
When we used this machine translation system with the Wikipedia dictionary, we found several entries that were inappropriate translations. For example, from the Wikipedia entry “フレンズ (Friends: TV drama from the US)”, dictionary entry “朋友” to “フレンズ (Friends)” was generated. However, because “朋友” is also used as a common term (i.e., an unnamed entity term), it is preferable to use “友人 (Friends)” instead of “フレンズ (Friends)” in such cases.
We checked manually the frequently occurring dictionary entries in a Chinese newspaper database to remove such inappropriate entries.
We also used Japanese redirection links to construct a synonym dictionary. All entries that had redirection links were normalized with respect to their destination entries. For example, using the redirection link “北朝鮮 (North Korea)” to “朝鮮民主主義人民共和国 (Democratic People’s Republic of Korea)”, “北朝鮮 (North Korea)” was normalized to “朝鮮民主主義人民共和国 (Democratic People’s Republic of Korea)” in the database.

In addition, because Wikipedia has a broad coverage of entries, from general topic terms to technical terms and from traditional terms to newly introduced terms, these entries may be good candidates for characteristic terms.

Based on this understanding, we developed a New NCon- strast system by modifying the NSContrast database to include an index of Wikipedia entries for each news article. We also restricted the candidates for term collocation analysis to Wikipedia entries.
The news article database was populated by following news sites from January 1, 2008 to April 19, 2009. Table 1 shows detailed information of the database.

Table 1: News Site Information

<table>
<thead>
<tr>
<th>Name of the site(country)</th>
<th>Number of articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL</td>
<td>Total/Daily</td>
</tr>
<tr>
<td>Asahi newspaper (Japan)</td>
<td>58344/122</td>
</tr>
<tr>
<td>Yomiuri newspaper (Japan)</td>
<td>48675/102</td>
</tr>
<tr>
<td>Nihon keizai newspaper (Japan)</td>
<td>69638/146</td>
</tr>
<tr>
<td>CNN (USA)</td>
<td>9542/20</td>
</tr>
<tr>
<td>Chosun newspaper (Korea)</td>
<td>24389/51</td>
</tr>
<tr>
<td>Joins newspaper (Korea)</td>
<td>18842/39</td>
</tr>
<tr>
<td>People newspaper (China)</td>
<td>18775/38</td>
</tr>
<tr>
<td>Chosun newspaper (Korea)</td>
<td>102009/214</td>
</tr>
<tr>
<td>Xinhua newspaper (China)</td>
<td>367459/773</td>
</tr>
</tbody>
</table>

Table 2 and Figure 2 shows an example of New NSCon- strast output when analyzing the term “Pakistan” during the first half of December, 2008. All terms were originally in Japanese and are translated into English.
In the table 2, we can see almost all countries shares same major topic keywords such as “Mumbai” and “India”. However keyword lists with Higher Change (minor topic with interest in each country) is different from other ones. In order to understand the relationships among these keywords, characteristic term collocation graph (Figure 2) is used. There are three types of node, in this term collocation graph. One is terms from Keyword list with higher correlation (white rectangle node) and the second is one with higher change (gray rectangle node). The last is name of the country (white octagon node). The locations of nodes and links are calculated in terms of the spring model [3] by using the GraphViz library with the fdp command.
For example, “Pacific Ocean” and “material” and “Obama” are Korean characteristic term and those keywords has strong collocation relationship with the keyword “nuclear weapon”. When the user clicking these keywords the system shows the news articles that contains these keywords. From these articles, the user can understand the Korean news paper pays more attention to the Obama’s policy related to the nuclear weapon issue about Pakistan and North Korea. In other words, the user can understand the Korean news paper pays more attention to the Obama’s policy related to the nuclear weapon issue about Pakistan and North Korea.
Table 2: List of Characteristic Keywords for analyzing Pakistan

<table>
<thead>
<tr>
<th>Country</th>
<th>Keyword list with Higher Correlation (major topic in each country or all countries)</th>
<th>Keyword list with Higher Change (minor topic with interest in each country)</th>
<th>Keyword list with Lower Change (comparatively neglected topic in each country)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>Mumbai, requiem (indot) Mistranslation of India</td>
<td>law, China, security, target</td>
<td>purpose, relation, western, broadcast</td>
</tr>
<tr>
<td>Japan</td>
<td>Mumbai</td>
<td>restriction Pacific Ocean, Palestine, improvement</td>
<td>dispatch, organization, director, president</td>
</tr>
<tr>
<td>Korea</td>
<td>India</td>
<td>integration material, boundary, commerce</td>
<td>anxiety, occurrence, security, unit</td>
</tr>
<tr>
<td>China</td>
<td>India</td>
<td>group, Obama base, usage</td>
<td>leader, terrorist, operation, border</td>
</tr>
<tr>
<td>USA</td>
<td>Mumbai</td>
<td>period, America soldier, morning</td>
<td>border, basement, China, dispatch</td>
</tr>
</tbody>
</table>

In this graph, keywords that represents common major topics are located in the middle of the graph. In this case “terrorist” “attack” at “Mumbai” “India” is a major topic in this period.

Comparing these results with those for the previous NSContrast system, the number of articles that have “Pakistan” as an index term has increased because of the synonym assimilation. In addition, the readability of the term collocation graph has improved.

4. CONCLUSION

In this paper, we have introduced the NSContrast system, which can extract characteristic information about news sites for given topic terms by using contrast set mining techniques. In addition, we have also proposed using Wikipedia to handle different expressions of the same named entity. By using Wikipedia in a New NSContrast system, we confirmed an improved recall of the news article retrieval results for a particular topic. This is important for an unbiased analysis of the differences between various countries.

5. REFERENCES


Figure 2: Term Collocation Graph for the Term “Pakistan”

Keywords related to Terrorism at Mumbai: India is a major topic of the world.

Korea pays attention to the news of nuclear weapon issue related to Pakistan. These news are also related to the Obama’s policy related to Pacific Ocean.
What’s wrong with Cross-Lingual IR?

John I. Tait
Information Retrieval Facility
Operngasse 20b
1040 Vienna, Austria
+43 1 236 94 74 6053
john.tait@ir-facility.org

ABSTRACT
To date much cross-language information retrieval research has focused on evaluation paradigms which were developed for monolingual web search. The paper argues that rather different scenarios are required for situations where cross-lingual search is a real requirement. In particular cross-lingual search is usually a collaborative as opposed to individual activity, and this needs to be taken into account in the evaluation of cross-lingual retrieval, especially when considering the notion of relevance.

Categories and Subject Descriptors H.3.3 [Information Search and Retrieval]

General Terms Documentation, Experimentation, Human Factors,

Keywords Patent Search; Intellectual property Search; Information Retrieval; Cross-lingual Retrieval.

1. INTRODUCTION
It seems to me that for non-professional searchers there is very little requirement for cross lingual searching. Most non-professional searchers formulate queries in their native language and require results in that language. Even with much better machine translation than has ever been available before one would rarely find an automatic translation that one could include in ones school homework!

On the other hand professional searchers in field like Intellectual Property, Competitor Analysis, opinion mining and some international aspects of legal search, for example really do need Cross Lingual Information Retrieval (CLIR). Other forms of professional search where CLIR is especially relevant include military intelligence analysis; employees of multi-national corporations where they are not fluent in major language of the corporation; and journalism.

Of course there are consumer searches where CLIR is needed: immigrants may require personal access to legal documents available only in the language of the country in which they are living.

This paper outlines one such professional information retrieval setting (patent search) and points out some problems with evaluation in that setting (especially the need for a sophisticated notion of relevance).

2. RELEVANCE IN CLIR
Experience with patent search has made it clear that while professional patent searchers need to access information in all languages in which patents can be filed: they require output in comparatively few languages: possibly only English and Chinese.

This has implications for the design of cross lingual information systems, but also the evaluation including the ways in which relevance is judged.

Now there are several steps in the judgement of relevance in this context. First, the searcher needs to make initial judgements of the relevance of patents (and indeed scientific articles and other material which may show the patent is not original, or obvious for instance). Then the patent attorney will review the results of the search; and in some cases other people: for example technical specialists (chemical engineers, molecular geneticists, search engine engineers etc.), language specialist, other lawyers, business managers and so on.

Now each of these groups, and the group collectively for an individual search, will bring different judgements of relevance to the retrieved document set. This needs to be taken into account and modelled explicitly in the evaluation.

Consider potential confounding factors in the experiment: what we are attempting to judge is the ability of the searcher to use the system to locate and determine the relevance (as assessed by the whole group). Quality of result translation may, for example, cause incorrect determination of relevance (or irrelevance) and we really need evaluation frameworks which take this into account.

It might be thought that evaluation of relevance after each stage would suffice: but the problem is that we wish to reuse the results of user centered experiments as a gold standard for subsequent system centered experiments. Subsequent group review of searchers initial relevance judgements is likely to cause changes in their view of what constitutes a relevant document, compounding the usual problems of having judged relevance sets which are consistent across time and judge.

I’m not claiming to say much new here: See Saracevic [3] for much more sophisticated approach: but those ideas do need to be more rigorously and consistently applied to CLIR evaluation.

3. OTHER ASPECTS OF EVALUATION
The consideration of confounding factors in our evaluation experiments leads onto some more general requirements of evaluations of CLIR for professional search. It is not appropriate to give an exhaustive list here, but factors to be taken into account include:
1. The place of the computer systems in the human system;
2. The need for component based evaluation;
3. The need to assess the impact of frozen and limited collections on the ecological validity of the experiment (for example where the collection excludes relevant documents searchers expect to see from their prior subject knowledge).

All this sounds more difficult thinking through than it has been done to date. Petrelli and others [4,5] have attempted to do so in the context of news but this work seems to have received comparatively little attention in subsequent literature.

Incidently, they do not fully report the language competence of their users, and their sample is small, but the papers seem to acting on behalf an inventor or their employer. More complex searches might be done at the behest of strategic business managers requesting patent landscape searching to determine, for example whether research and development investment in a particular area is likely to yield patentable results.

The patent searcher will then formulate a series of queries (the search strategy) which will be addressed to often several different search systems. In practice most searching is on English abstracts, but really thorough searching for patentability for example requires searching of many different languages. This is a high recall task, in which it is important not to miss relevant documents.

Now there are several steps in the judgement of relevance in this context. First, the searcher needs to make initial judgements of the relevance of patents (and indeed scientific articles and other material which may show the patent is not original, or obvious for instance). Then the patent attorney will review the results of the search; and in some cases other people: for example technical specialists (chemical engineers, molecular geneticists, search engine engineers etc.), language specialist, other lawyers, business managers and so on.

Now each of these groups, and the group collectively for an individual search, will bring different judgements of relevance to the retrieved document set. This needs to be taken into account and modelled explicitly in the evaluation.

Consider potential confounding factors in the experiment: what we are attempting to judge is the ability of the searcher to use the system to locate and determine the relevance (as assessed by the whole group). Quality of result translation may, for example, cause incorrect determination of relevance (or irrelevance) and we really need evaluation frameworks which take this into account.

It might be thought that evaluation of relevance after each stage would suffice: but the problem is that we wish to reuse the results of user centered experiments as a gold standard for subsequent system centered experiments. Subsequent group review of searchers initial relevance judgements is likely to cause changes in their view of what constitutes a relevant document, compounding the usual problems of having judged relevance sets which are consistent across time and judge.

Now I’m not claiming to say much new here: See Saracevic [3] for much more sophisticated approach: but those ideas do need to be more rigorously and consistently applied to CLIR evaluation.

The consideration of confounding factors in our evaluation experiments leads onto some more general requirements of evaluations of CLIR for professional search. It is not appropriate to give an exhaustive list here, but factors to be taken into account include:
1. The place of the computer systems in the human system;
2. The need for component based evaluation;
3. The need to assess the impact of frozen and limited collections on the ecological validity of the experiment (for example where the collection excludes relevant documents searchers expect to see from their prior subject knowledge).

All this sounds more difficult thinking through than it has been done to date. Petrelli and others [4,5] have attempted to do so in the context of news but this work seems to have received comparatively little attention in subsequent literature.

Incidently, they do not fully report the language competence of their users, and their sample is small, but the papers seem to acting on behalf an inventor or their employer. More complex searches might be done at the behest of strategic business managers requesting patent landscape searching to determine, for example whether research and development investment in a particular area is likely to yield patentable results.

The patent searcher will then formulate a series of queries (the search strategy) which will be addressed to often several different search systems. In practice most searching is on English abstracts, but really thorough searching for patentability for example requires searching of many different languages. This is a high recall task, in which it is important not to miss relevant documents.

Now there are several steps in the judgement of relevance in this context. First, the searcher needs to make initial judgements of the relevance of patents (and indeed scientific articles and other material which may show the patent is not original, or obvious for instance). Then the patent attorney will review the results of the search; and in some cases other people: for example technical specialists (chemical engineers, molecular geneticists, search engine engineers etc.), language specialist, other lawyers, business managers and so on.

Now each of these groups, and the group collectively for an individual search, will bring different judgements of relevance to the retrieved document set. This needs to be taken into account and modelled explicitly in the evaluation.

Consider potential confounding factors in the experiment: what we are attempting to judge is the ability of the searcher to use the system to locate and determine the relevance (as assessed by the whole group). Quality of result translation may, for example, cause incorrect determination of relevance (or irrelevance) and we really need evaluation frameworks which take this into account.

It might be thought that evaluation of relevance after each stage would suffice: but the problem is that we wish to reuse the results of user centered experiments as a gold standard for subsequent system centered experiments. Subsequent group review of searchers initial relevance judgements is likely to cause changes in their view of what constitutes a relevant document, compounding the usual problems of having judged relevance sets which are consistent across time and judge.

Now I’m not claiming to say much new here: See Saracevic [3] for much more sophisticated approach: but those ideas do need to be more rigorously and consistently applied to CLIR evaluation.

The consideration of confounding factors in our evaluation experiments leads onto some more general requirements of evaluations of CLIR for professional search. It is not appropriate to give an exhaustive list here, but factors to be taken into account include:
1. The place of the computer systems in the human system;
2. The need for component based evaluation;
3. The need to assess the impact of frozen and limited collections on the ecological validity of the experiment (for example where the collection excludes relevant documents searchers expect to see from their prior subject knowledge).

All this sounds more difficult thinking through than it has been done to date. Petrelli and others [4,5] have attempted to do so in the context of news but this work seems to have received comparatively little attention in subsequent literature.

Incidently, they do not fully report the language competence of their users, and their sample is small, but the papers seem to
indicate a predominant need (in CLIR) to retrieve non-English documents for (manual or automatic) translation into English.

What is really required, as a first step, is the ability to observe the use of patent search in situ with sufficient access to all those involved in patent searching to record and explore the reasons why they made the relevance judgements they did, and to at least assess the likelihood those judgements will be consistent and reproducible. There are considerable barriers to doing this: many users of the patent system are extremely secretive because their patent searches can reveal information about their commercial and research strategies which is of immense value to their competitors. However if we as scientists can achieve the necessary level of trust among these communities it will both yield value dat for us and at the very least better assessments for them of their own working practices.

4. CONCLUSION

Conventional CLIR evaluations have relied very much on the Cranfield experimental model pioneered by Cyril Claverdon, Kaern Sparck Jones and others [6]. This paper is really a plea to move to more sophisticated models of evaluation for professional search, the context in which cross lingual retrieval is really valuable.

5. ACKNOWLEDGMENTS

I would like to thank my colleagues on the evaluation working group at the recent Interactive Information Retrieval Dagstuhl who developed my thinking on this topic; my colleagues in Matrixware and the IRF especially those who have worked on the CLEF IP and TREC CHEM tracks; the many IP professionals who have taken the time to educate me about patent search – especially Henk Tomas; my three my anonymous reviewers and the workshop organizers for some valuable criticisms and suggestions.

6. REFERENCES

A Virtual Evaluation Track for Cross Language Link Discovery

Wei Che (Darren) Huang  
Faculty of Science and Technology  
Queensland University of Technology  
Brisbane, Australia  
w2.huang@student.qut.edu.au

Andrew Trotman  
Department of Computer Science  
University of Otago  
Dunedin, New Zealand  
andrew@cs.otago.ac.nz

Shlomo Geva  
Faculty of Science and Technology  
Queensland University of Technology  
Brisbane, Australia  
s.geva@qut.edu.au

ABSTRACT
The Wikipedia has become the most popular online source of encyclopedic information. The English Wikipedia collection, as well as some other languages collections, is extensively linked. However, as a multilingual collection the Wikipedia is only very weakly linked. There are few cross-language links or cross-dialect links (see, for example, Chinese dialects). In order to link the multilingual-Wikipedia as a single collection, automated cross language link discovery systems are needed – systems that identify anchor-texts in one language and targets in another. The evaluation of Link Discovery approaches within the English version of the Wikipedia has been examined in the INEX Link-the-Wiki track since 2007, whilst both CLEF and NTCIR emphasized the investigation and the evaluation of cross-language information retrieval. In this position paper we propose a new evaluation track: Cross Language Link Discovery (CLLD). The track will initially examine cross language linking of Wikipedia articles. This virtual track will not be tied to any one forum; instead we hope it can be connected to each of (at least): CLEF, NTCIR, and INEX as it will cover ground currently studied by each. The aim is to establish a virtual evaluation environment supporting continuous assessment and evaluation, and a forum for the exchange of research ideas. It will be free from the difficulties of scheduling and synchronizing groups of collaborating researchers and alleviate the necessity to travel across the globe in order to share knowledge. We aim to electronically publish peer-reviewed publications arising from CLLD in a similar fashion: online, with open access, and without fixed submission deadlines.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – Search Process.

General Terms: Measurement, Performance, Experimentation

Keywords: Cross Language, Link Discovery, Information Retrieval, Evaluation

1. INTRODUCTION

1.1 Background
Collaborative hypertext knowledge management systems, such as the Wikipedia, offer an efficient means for creating, maintaining, and sharing information. Extensive linking between documents in these systems is essential for user navigation and assists readers to varying degrees. A reader with extensive background knowledge of a topic may be less likely to follow a link, while a less knowledgeable reader may choose to follow many links in order to expand their knowledge.

Links in the Wikipedia originate from two primary sources: the page authors’ knowledge of the document collection; and automated link discovery systems such as We Can Link It [6] and Wikify [9]. The Link-the-Wiki track at INEX [22] was established as an independent evaluation forum for measuring the performance of these kinds of link discovery systems.

In 2007 the track explored document-to-document link discovery in the English Wikipedia, in 2008 the track also looked at anchor-text identification within the source document and the placement of the target point (the best-entry-point, or BEP) within the target document. This second kind of link discovery is known as anchor-to-BEP link discovery, or focused link discovery. The track also developed a standard methodology and metrics for the evaluation of link discovery systems.

The track’s results show that excellent link discovery systems have been developed – that is, there are now published algorithms that can almost perfectly predict the links in a Wikipedia page. However, manual assessment revealed the highly unexpected result that many existing Wikipedia links are not relevant (at least not to the INEX assessors)! INEX now recommends manual assessment as the preferred procedure for the evaluation of link discovery systems. We note that an INEX assessor manually assessing links from the pool perfectly models a user who (after adding a new article to the Wikipedia) is navigating a list of links recommended by a link discovery system – accepting or rejecting as they go. This process lends itself to the interactive study of link discovery systems.

With the growth of the multilingual Wikipedia (and the multilingual web) there is a growing need for cross-language information retrieval including cross language interlinking of multilingual documents. Most Wikipedia pages are written in English and we, unsurprisingly but anecdotally, observe users whose first language is not English searching the English Wikipedia. Their need is two-fold: the Wikipedia documents to be translated into their first language; and links between documents...
to reflect their language choices. Translation is already happening and some cross-language links already exist, however these problems are our research motivation. We are trying to:

- Identify Wikipedia documents that are all on the same topic, irrespective of language; and
- Identify hypertext links between documents, irrespective of language, so that a user can choose a target document based on a language preference.

1.2 Motivation

Many Internet users are multi-lingual. To satisfy their information needs the search engine should return documents in the different languages they read. Doing so is more thorough than returning results in just one language. As examples, the Early history of the United States of America can be found in the Chinese Wikipedia but the English Wikipedia has a much richer document on the topic; information about Chinese Dynasties may be found in several documents in the Chinese language Wikipedia, and indeed in several distinct Chinese language version of the Wikipedia. In both examples, a link between these different language versions will help the multi-lingual reader. In both examples focused cross-lingual anchor-to-BEP links would result in a more comprehensive interlinked knowledge base, especially if the links the multilingual reader sees are based on a personal language profile. Envisage a document being interlinked to any number of languages, but users only seeing links to languages that are defined in their personal profile.

Anchor-to-BEP linking is a feature of HTML that is rarely exploited in links – despite its existence since the beginnings of the web. Very few links in the Wikipedia actually take the user from the point of reference (the anchor) to the target location within another document (the BEP). Such interlinking is common within a single document and is used in navigation, but is rarely utilised when linking between documents. Such focused interlinking is particularly desirable when documents are large or when browsing on small mobile devices. For instance, in the article South Eastern Main Line, an orphaned anchor, Folkestone Harbour, is colored in red. It is a place-holder for a link to an article that does not yet exist. However, the article Folkestone does have a section titled Folkestone Harbour. This prospective anchor could be linked to this section until an article on Folkestone Harbour is created.

Extending the INEX Link-the-Wiki track to cross language linking will help turn the Wikipedia into a multi-lingual knowledge network. The section 地理 (English: Geography) in the article 多佛 (English: Dover Harbour) and 多佛海底隧道 (English: Channel Tunnel; French: Le tunnel sous la Manche). There is no link for 多佛 (English: Dover Harbour) in the Chinese Wikipedia, but an article on the Port of Dover is linked from the redirect of the Dover Harbour page in the English Wikipedia. Information (images and geography) about Dover Harbour can also be found in the English article on Dover. The article, 多佛海底隧道, does not express much information about the channel tunnel, certainly not as much as the English Channel Tunnel page. These two examples show the need for cross-language links within the Wikipedia.

To the best of our knowledge, current link discovery systems such as Wikify [9] focus on monolingual Wikipedia and have not been extended to support multilingual link discovery. Cross-language tracks conducted in NTCIR and CLEF explore Information Retrieval and Question-Answering but not link discovery. Link Discovery is different from Information Retrieval although it does rely on similar technology: for link discovery a match of semantic context between the point of reference (the context of the anchor) and the target text (the BEP context) is essential. Unlike query based information retrieval, in link discovery the context of the anchor is always explicit since the anchor is always embedded in surrounding text, and similarly the context of the target location is highly focused and specific. In information retrieval the query is known but the context unknown, in link discovery it is necessary to identify both the query (the anchor-text) and the results list (the target document and BEP) and embedding contexts are available.

Link discovery provides a rich context in which NLP based approaches may well prove much more useful than they had been in the query based information retrieval. Furthermore, cross-language link discovery involves a set of technologies, including IR, NLP, semantic and similarity matching techniques, character encoding technologies, machine readable corpora and dictionaries, machine translation, focused and passage retrieval, and multiple links per anchor discovery. Cross Language Link Discovery (CLLD) demands the tight integration of techniques currently under examination at INEX, CLEF and NTCIR.

Herein we formally propose the CLLD track. This track will be run as a single collaborative web-based forum. Participants will be drawn from the existing forums, but be part of none (or all), it will be an online virtual evaluation forum. All collections, topics, submission and result analysis will be maintained via a remote repository. By using only public domain data (such as the Wikipedia and open source software) we can simplify participation and the sharing of resources. The community of participants will provide software tools and assessments; as well as a peer-reviewed online publication for approaches and results. The forum will not be tied to any particular timeline or venue but will be run as a continuous evaluation track – without a dedicated annual event (although there is no reason not to hold such meetings, perhaps as surrogate to larger events). This proposal represents a dramatic philosophical change from the traditional TREC paradigm.

2. RELATED WORK

As suggested by Wilkinson & Smeaton [1], navigation between linked documents is a great deal more than simply navigating multiple results of a single search query, linking between digital resources is becoming an ever more important way to find information. Through hypertext navigation, users can easily understand context and realize the relationships in related information. However, since digital resources are distributed it has become difficult for users to maintain the quality and the consistency of links. Automatic techniques to detect the semantic structure (e.g. hierarchy) of the document collection, and the relatedness and relationships of digital objects have been studied and developed [2]. Early works, in the 1990s, determined whether and when to insert links between documents by computing document similarity. Approaches such as term repetition, lexical chains, keyword weighting and so on were used to calculate the similarity between documents [3, 4, 5]. These approaches were
focused on the document-to-document linking scenario, rather than identifying which parts of which documents were related. Jenkins [6] developed a link suggestion tool, *Can We Link It*. This tool extracts a number of anchors which have not been discovered in the current article and that might be linked to other Wikipedia documents. The user can accept, reject, or click “*don’t know*” to leave a link as undecided. Using this tool the user can add new anchors and corresponding links back to a Wikipedia article.

A collaborative knowledge management system, called *PlanetMath*, based on the Noosphere system has been developed for mathematics [7]. It is encyclopedic, (like the Wikipedia), but mainly used for the sharing of mathematical knowledge. Since the content is considered to be a semantic network, entries should be cross-referenced (linked). An automatic linking system provided by Noosphere employs the concept of conceptual dependency to identify each entry for linking. Based on the Noosphere system, *NNexus* (Noosphere Networked Entry eXtension and Unification System) was developed to automate the process of the automatic linking procedure [8]. This was the first automatic linking system which eliminates the linking efforts required by page authors.

The *Wikipedia* [9] system which integrates technologies of automatic keyword extraction and word sense disambiguation can identify the important concepts in a document and link these concepts to corresponding documents in the Wikipedia. Mihalcea and Csomai stated that many of applications such as the annotation of corresponding documents in the Wikipedia. Mihalcea and Csomai stated that many of applications such as the annotation of corresponding documents in the Wikipedia. Mihalcea and Csomai stated that many of applications such as the annotation of corresponding documents in the Wikipedia. Mihalcea and Csomai stated that many of applications such as the annotation of corresponding documents in the Wikipedia.

Since the inception of TREC in 1992 interest in IR evaluation has increased rapidly and today there are numerous active and popular evaluation forums. It is now possible to evaluate a diverse range of information retrieval methods including: ad-hoc retrieval, passage retrieval, XML retrieval, multimedia retrieval, question answering, cross language retrieval, link discovery, learning to rank, and so on.

The CLIR (Cross-Language Information Retrieval) track was first introduced to TREC in 2000. It offered document collections in English, French and German and queries in English, French, German, Spanish and Dutch. Three fundamental resources, machine translation, machine readable dictionaries and corpus-based resources, have been used. There are three common approaches to match queries with the resource documents [10]: machine translation technology using dictionaries and statistical information or example-based translation; machine readable bilingual dictionaries; and relying on corpus resources to train retrieval mechanism by using Latent Semantic Indexing (LSI), Generalized Vector Space Model or Similarity Thesaurus for translating queries. As a performance baseline corresponding language queries were also submitted against the same language collections.

The cross-language track investigated the retrieval of relevant documents that pertain to a particular topic or query regardless of the language in which both the topic and documents were written. The last TREC CLIR track was run in 2002, whoever ongoing effort can be found in both NTCIR and CLEF [11].

NTCIR started in late 1997 and is focused on Japanese and other East Asian languages [12]. The workshop runs on an 18-month cycle. The aim is to build an infrastructure for large-scale experimental evaluation of Information Access (IA) research. IA in the workshop has been indicated as the process of searching, browsing and looking for relevant information, and utilizing the information. Technologies, like Information Retrieval (IR), Cross-Language Information Retrieval (CLIR), Question-Answering (QA), text summarization and text mining, are considered part of the IA family.

The goal is to develop a module-based infrastructure for evaluation integrating IR and QA technologies to propose answers in a suitable format for given questions in any language. It intends to model significant work from every participant and build a set of APIs (or modules) to facilitate the development of cross-language (QA) systems. A platform, called EPAN (Evaluation Platform for_ACLIA and NTCIR), was adopted by NTCIR to perform the collaborative evaluation [13]. Through module-by-module evaluation it is possible to identify problems in parts of participants’ otherwise complicated CLIR-QA systems – something not possible in end-to-end evaluation. For example, many CLIR-QA systems failed to retrieve relevant documents when named entities in queries did not appear in ordinary translation dictionaries. The module-based approach also makes it possible for participants to collaborate by working on different modules.

The annual CLEF forum aims to create a research forum in the field of multilingual system development [14]. The experiments range from monolingual text retrieval to multilingual multimedia retrieval. The collection and available languages vary depending on different tasks. They include 3 million News articles in 13 languages, a social science database in English and German, the Cambridge Sociological Abstracts, and the Russian ISIS collection, 3.5 million web pages in multi-languages, and a photograph database with captions in different languages [15]. Various sets of topics in different languages are available for respective tasks.

The DIRECT system used by CLEF manages data (collection, topics, and metrics), builds statistics (analysis, plots and results) and provides different entries for various roles [16, 17]. Of particular note is the dynamic user interface through which participants can interact with their-own and others’ experimental data and results.

3. MULTILINGUAL WIKIPEDIAS

The Wikipedia is a multilingual online encyclopedia that offers a free and flexible platform for developing a collaborative knowledge repository [18]. Currently, it has entries written in more than 200 different languages. Overell [19] shows that the geographic coverage of the Wikipedia very much depends on the language version – places in the UK are best covered by the English language version of the Wikipedia while places in Spain are best covered by the Spanish language version. There are more than 2,874,919 articles (May 2009) in the English Wikipedia which is the largest language version in the Collection. By the end of 2008, no fewer than 13 languages have more than 200,000 articles. As can be seen in Table 1, both European and Asian language versions have reached a substantially size – the collection is already useful for multilingual research [20].
Most documents have a rich set of same-language links but a much lower number cross-language links. Figure 1 shows that by October 2006 (Wikipedia have not provided stats for English since that date) only 1.4 million English anchor texts have been linked to entries in other languages while the amount of same-language links in the English Wikipedia had reached 24.2 million.

Around a third of anchor texts in the Chinese Wikipedia were linked to other language collections in the Wikipedia. There are several different Chinese dialects available as different collections and many of English terms are linked to the English Wikipedia. Other versions shown in Figure 1 have less than 20 percent of their links pointing to other languages. For the Article-to-Article, Sorg and Cimano state that around 50% of Wikipedia articles in German version are linked to the English Wikipedia articles whilst only 14% of English articles have links to the German version [21]. Linking the Wikipedia entries across different languages is still very limited. Efficient and accurate cross-language link discovery is yet to be demonstrated and evaluated.

<table>
<thead>
<tr>
<th>Language</th>
<th>Articles (K)</th>
<th>Database (GB)</th>
<th>Bytes per Article</th>
<th>Over ZKB</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>1100</td>
<td>3.2</td>
<td>3092</td>
<td>365 (33%)</td>
</tr>
<tr>
<td>German</td>
<td>858</td>
<td>3.5</td>
<td>3489</td>
<td>420 (49%)</td>
</tr>
<tr>
<td>French</td>
<td>746</td>
<td>2.7</td>
<td>2995</td>
<td>269 (36%)</td>
</tr>
<tr>
<td>Polish</td>
<td>567</td>
<td>1.5</td>
<td>1988</td>
<td>130 (23%)</td>
</tr>
<tr>
<td>Japanese</td>
<td>555</td>
<td>2.6</td>
<td>1890</td>
<td>128 (23%)</td>
</tr>
<tr>
<td>Italian</td>
<td>533</td>
<td>1.9</td>
<td>2914</td>
<td>192 (36%)</td>
</tr>
<tr>
<td>Dutch</td>
<td>507</td>
<td>1.3</td>
<td>2921</td>
<td>147 (29%)</td>
</tr>
<tr>
<td>Portuguese</td>
<td>448</td>
<td>1.1</td>
<td>1866</td>
<td>90 (20%)</td>
</tr>
<tr>
<td>Spanish</td>
<td>430</td>
<td>1.8</td>
<td>3492</td>
<td>189 (44%)</td>
</tr>
<tr>
<td>Russian</td>
<td>347</td>
<td>1.3</td>
<td>2822</td>
<td>125 (36%)</td>
</tr>
<tr>
<td>Swedish</td>
<td>301</td>
<td>0.609</td>
<td>1535</td>
<td>54 (18%)</td>
</tr>
<tr>
<td>Chinese</td>
<td>209</td>
<td>0.717</td>
<td>1451</td>
<td>36 (17%)</td>
</tr>
<tr>
<td>Norwegian</td>
<td>203</td>
<td>0.475</td>
<td>1781</td>
<td>43 (21%)</td>
</tr>
</tbody>
</table>

* The statistics of the English version was dated Oct. 2006.

Figure 1: Different linking types in different language versions of the Wikipedia at the end of 2008

In order to achieve a comprehensive cross-language knowledge network, cross-language link discovery is essential. The document collection is large (there are many millions of documents) and broad (there are many languages covered). Although only one link per anchor is typically displayed by existing HTML based web browsers, there is no inherent restriction for this limit in the HTML standard. Anchor text can be linked not only to multiple targets, but also in different languages, and the extension of browsers to support this functionality is long overdue.

4. INEX LINK-the-WIKI TRACK

The INEX Link-the-Wiki track offers a standard forum for the evaluation of link discovery in both document-to-document and anchor-to-BEP linking. The task is to discover a set of anchor texts, based on the context of the topic, which can semantically be linked to best entry points in other documents. Besides outgoing links, candidate incoming links from other articles into the topic document are also required.

The INEX 2008 English Wikipedia collection consisting of 659,388 documents was used as the corpus for the experiments since Wikipedia is composed of millions of interlinked articles in numerous languages and have been proved as a valuable collection for IR research. 50 documents nominated by participants were used in the anchor-to-BEP task. For the document-level link discovery, 6,600 documents were randomly selected but pre-filtered (for suitability) by size and the number of links. For 2009 this collection has been updated and now consists of over 2.6 million articles, spanning over 60GB of text (without images).

The documents were converted into topics by removing all links in the documents and removing all links from the remainder of the collection to those documents, they were said to have been orphaned from the collection. The original documents (with the links) were said to be pre-orphans.

Links within the collection to and from the pre-orphans were extracted and used as the ground truth to which runs were compared. This is the automatic evaluation method. Using standard methodology, the topics were sent to participating groups. Participant’s link discovery systems were required to return a ranked list of at most 50 outgoing text anchors (each of which targeted up-to 5 documents). The results were exhaustively pooled and the ground truth of existing links in the Wikipedia was added to the pools. Runs were manually assessed to completion, by the groups that nominated the topics. This formed the manual evaluation set. Assessment against the ground-truth result set generated a score for the performance of a submission relative to the Wikipedia. Additional evaluation was performed against the manually assessed links (including the manually assessed existing Wikipedia links). The outcome showed that the Wikipedia ground-truth does not agree with user expectation. The manual assessment process is necessary in order to produce a reliable test base for evaluation [22].

5. PROPOSED CLLD METHODOLOGY

The cross language link discovery track will work in a similar manner to the INEX track. Participating groups will be asked to submit a list of languages they can read (from those covered by the Wikipedia). This also indicates the languages these participants can submit runs for and can assess in. A subset of topics from the different language Wikipedia collections will be chosen and distributed to participants. To support the creation of topics, a selection tool will be developed to help choose and to orphan documents in indicated languages. Participants will be able to choose any combination of languages, for example
German documents linking to English or Dutch documents, or Chinese documents linking to Japanese documents.

Programs will be provided that allow participants to view their runs. These programs will show proposed linked documents with the anchors and their respective target document best-entry-points, as would be seen by an assessors (and ultimately end-users of their system).

The challenge for the organizers is to obtain a critical mass of participants and assessors to facilitate robust and reliable manual evaluation in multiple languages. The track must, therefore, be a close and extensive collaboration between NTCIR, CLEF, INEX, and other evaluation forums.

5.1 Tasks Specification
Initially two linking tasks will be formalized:

- **MULTILINGUAL** topical linking:
  This is a form of document clustering — the aim is to identify (regardless of language) all the documents in all languages that are on the same topic. The Wikipedia currently shows these links in a box on the left hand side of the page.

- **BILINGUAL** anchor linking:
  It is exemplified by the Chinese article 诺森伯兰郡, having a link from the anchor 邊境县 to the English article List of United Kingdom Parliament constituencies. The link discovery system must identify the anchor text in one language version of the Wikipedia and the destination article within any other language version of the Wikipedia.

In the case of MULTILINGUAL topical linking, the participants are encouraged to discover as many documents as they can.

In the case of BILINGUAL anchor linking, at most 50 anchors may be identified in a orphaned document and up to 5 BEPs may be linked to one target language (e.g. English to German). Initially only outgoing links will be examined since incoming links from a single language may not make sense.

5.2 Test Collections and Submission
The set of multilingual Wikipedia collections will be used as the corpus for cross language link discovery. The size and the number of documents are listed in Table 1. Nominated topics will be collected and the ground-truth extracted from the collection.

Participants will be encouraged to share in the development of appropriate procedures for topic selection, multilingual topic discovery, ground truth link extraction, and the assessment method.

The submission format may be derived from the format currently used by INEX. The existing INEX tools will be ported to support CLLD.

5.3 Evaluation
It is essential to define a standard methodology and procedure to assess link quality and to quantitatively evaluate different approaches.

5.3.1 Static Evaluation
When Trotman & Geva [24] introduced the Link-the-Wiki track at INEX they also noted that the evaluation required no human assessment. The same is true with cross-language link discovery.

Topics in the INEX Link-the-Wiki track are chosen directly from the document collection. All links in those documents are removed (the document is orphaned). The task is to identify links for the orphan (both from and to the collection). Performance is measured relative to the pre-orphan (the document before the links were removed).

For **MULTILINGUAL** linking the links on the left hand side of the Wikipedia page could be used as the ground truth. The performance could be measured relative to the alternate language versions of the page already known to exist.

BILINGUAL anchor linking from one document to another could also be automatically evaluated. Links from the pre-orphan to a destination page in an alternate language would be used as the ground truth – but there are unlikely to be many such links.

A same-language link from a pre-orphan to a target provides circumstantial evidence that should the target exist in multiple languages then the alternate language versions are relevant. This is essentially a triangulation:

\[ A \rightarrow B \rightarrow C \Rightarrow A \rightarrow C \]

where \( A, B, \) and \( C \) are articles; \( t \) designates a topical link, \( l \) a cross language link, and \( r \) a topical cross language link. By extension, if \( A \rightarrow B \rightarrow C \Rightarrow A \rightarrow C \), where \( B \rightarrow C \) designated two documents that are not linked but have the same title.

Static assessment requires no human interaction. A web site with orphan sets chosen using some criteria (perhaps randomly), with the assessment sets (extracted from the pre-orphans), and that will evaluate a run will be built and provided. Such an evaluation methodology raises the possibility of running the track continually and without any deadlines.

Huang et al. [22] question automatic evaluation. Their investigation suggests that many of the links in the Wikipedia are not topical, but are trivial (such as dates), and that users do not find them useful. Manual assessment is, consequently, necessary. Although the automatic evaluation method in the INEX Link-the-Wiki is less accurate, it is still very practical and provides a reliable way to evaluate different link discovery methods against the ground truth. The submission runs that have better performance in automatic evaluation tend to have better results in manual evaluation.

5.3.2 Continual Evaluation
Manual assessment raises new challenges for cross language link discovery because finding assessors fluent in multiple languages is difficult – especially for a track with a relatively small number of participants but in a large number of languages (the Wikipedia has over 200 languages).

We propose a novel form of evaluation called *continual evaluation* in which participants can download topics and submit runs at any time; and in which contribution to manual assessment is an on-going concern. The document collection will, initially, be static. Topics will either be chosen at random from the collection, or nominated by participants. For any given run a participant will download a selection of topics and submit a run. The evaluation will be based on metrics that consider the un-assessed document problem (such as a variant on rank-biased precision [23]), and comparative analysis will be relative to an incomplete, but growing, assessment set.
To collect assessments two methods are proposed: first, in order to submit a run the participant will be required to assess some anchor-target pairs in languages familiar to them; second, we will run an assessment Game With A Purpose (GWAP). Kazai et al. used a GWAP for the INEX Book track; Von Ahn & Dabbish [25] discuss GWAPS in other contexts (including the Google Image Labeler). Regardless of the method of assessment collection, we are trying to validate the minimum number of links necessary to disambiguate the relative rank order of the runs (within some known error).

6. PUBLICATION
Both automatic and manual assessment of cross language link discovery can be done on a continual rolling basis; there is no need for topic submission deadlines, run deadlines, assessment deadlines; or paper publication deadlines.

At INEX the time difference between run-submission and the workshop paper submission date is long (6 July – 23 Nov). With automatic assessment it is possible to achieve a result, write, and then publish a paper with a short turn around. As part of the virtual track we propose an open-access virtual workshop workbook to which registered participants can immediately submit their papers for peer-review and publication.

7. CONCLUSIONS AND OUTLOOK
As far as we are aware, the cross-language link discovery track is the first to offer extensive reusable independent evaluation resources. In this paper we introduce this new evaluation task.

A fully automated procedure for anchor-to-document link analysis, using the existing Wikipedia linking network is described. The procedure was used at INEX 2007 and allowed us to create a fast evaluation procedure with a turnaround time of days and not months because it had no manual assessment. The procedure allows for a very large number of documents to be used in experiments. This overcomes the assessment bottleneck which is encountered in most other tasks in collaborative evaluation forums such as INEX and TREC. We further proposed to extend the task to Cross Language Link Discovery, and propose the concept of automatic evaluation. We describe the requirement for evaluating such a task.

These activities may not be held in a fixed place but can be done by gathering participants from INEX, CLEF and NT CIR through a virtual web-based system. The CLLD track will be dedicated to supporting efficient methods and tools for CLLD evaluation. The collections, submission and result data will be well managed for further analysis and experiments. Participants from different nations are expected to work collaboratively to achieve the development of multilingual link discovery systems.

Baseline automatic evaluation methods seen at INEX do not require human intervention as the assessments are extracted directly from the collection and performance is measured relative to these. The new track can, therefore, bootstrap and run online with continuous evaluation, free from the problems of scheduling groups of collaborating researchers. Overtime manual assessments will be collected and improve the available resources. We also propose to publish the results of the track in a similar fashion to the CLLD track itself – online, with open access, and with quality control.

8. REFERENCES


Towards an Integrative Approach to Cross-Language Information Access for Digital Libraries

Jiangping Chen
Department of Library and Information Sciences,
College of Information, University of North Texas
1155 Union Circle 311068, Denton, Texas 76203-1068, USA
Jiangping.Chen@unt.edu

Miguel E Ruiz
Department of Library and Information Sciences,
College of Information, University of North Texas
1155 Union Circle 311068, Denton, Texas 76203-1068, USA
Miguel.Ruiz@unt.edu

ABSTRACT
Digital libraries are becoming more ubiquitous and their contents can be of interest not only to the communities that create them but also to individuals and organizations around the world. The content of most digital libraries is not visible to web search engines and the internal search engines of many digital libraries do not support cross language access. In this paper, we present some ideas for a general framework to support cross-language information access for digital libraries. The Integrative Cross Language Information Access (iCLIA) framework is designed to facilitate cross language access to digital libraries by capitalizing on the integration of multiple sources for translational knowledge, integration of computational power and human intelligence, and integration of several methods for words sense disambiguation. Collaboration among digital library software developers, CLIR researchers and users is extremely important for the application of CLIR to existing and future digital library projects.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.3.7 [Information Storage and Retrieval]: Digital Libraries.

General Terms
Information Retrieval, Digital Libraries, Standardization

Keywords

1. INTRODUCTION
Digital libraries contain human intellectual creations that are valuable not only to designated user communities, but also to interested organizations and individuals all over the world. The content of most digital libraries is not visible to web search engines; therefore sophisticated digital library (DL) services such as searching and browsing have been developed and explored to facilitate information access. However, most existing digital libraries can only be accessed in a single language. Cross-Language Information Access (CLIA), an extension of the field of Cross-Language Information Retrieval (CLIR) [5], enables DL users to search for and locate digital objects written in languages other than the language of their search terms. Although CLIA technologies including CLIR, Cross-language Question Answering (CLQA), and Cross-Language Information Extraction (CLIE) have been actively explored by researchers since mid-1990s, none of the technologies have been widely applied to existing digital libraries to enable digital library users to access information across languages [4, 7].

Due to the fact that machine translation usually produces hard-to-understand translations, many organizations and information systems still rely on human translators for translating documents or files from one language to another. As for digital libraries, very few have realized multilingual information access [2]. An analysis of 150 digital libraries in the United States found out that only five of them can be accessed by using more than one language. Table 1 lists these five digital libraries.

Table 1. Digital Libraries with Multilingual Information Access in the United States.

<table>
<thead>
<tr>
<th>Library Name</th>
<th>URL</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meeting of Frontiers</td>
<td><a href="http://frontiers.loc.gov/intldl/mtfhtml/mfap">http://frontiers.loc.gov/intldl/mtfhtml/mfap</a> slash.html</td>
<td>English/Russian</td>
</tr>
<tr>
<td>France in America</td>
<td><a href="http://international.loc.gov/intldl/fiahtml/fi">http://international.loc.gov/intldl/fiahtml/fi</a> ahome.html</td>
<td>English/French</td>
</tr>
<tr>
<td>Parallel Histories</td>
<td><a href="http://international.loc.gov/intldl/eshhtml/">http://international.loc.gov/intldl/eshhtml/</a></td>
<td>English/ Spanish</td>
</tr>
<tr>
<td>The Perseus Digital Library</td>
<td><a href="http://www.perseus.t">http://www.perseus.t</a> ufts.edu</td>
<td>Greek, English, Latin</td>
</tr>
</tbody>
</table>

The digital libraries listed in Table 1 have been funded by various funding agencies, especially the federal government.
They are each the product of collaboration. DL developers in different countries cooperate to produce the bilingual or multilingual collections. These digital libraries serve broad or global user communities in which users speak different languages. However, none of them employ cross-language search, cross-language information retrieval techniques or machine translation. This situation may be the result of concerns involving technical, financial, and user related issues. However, a lack of exploration of CLIA framework for digital libraries and practical assessment of such framework may prevent digital library communities from implementing CLIA services.

The European Union has been a strong supporter of the creation of digital libraries that support multilingual access. The European Library (TEL) is one of the projects that has shown interest in supporting cross-lingual access to the collections of National Libraries members of the EU [3], but so far it only allows monolingual retrieval in many of the different languages and multilingual access through matching of the controlled vocabulary assigned as part of the metadata. According to Cousins [3] TEL conducted a survey and found that users have high expectations of the integrated portal, expecting it to be able to support cross-lingual retrieval access. Another recent example of an attempt to create a library with CLIA support is the Europeana project 1, which supports multilingual access in 26 European Languages. Other examples include the Cross-Language Access to Catalogues and On-Line Libraries (CACAO) 2; Improving Access to Text (IMPACT) 3, and the Common Language Resources and Technology Infrastructure (CLARIN) 4.

Cross language information retrieval for web search engines have been made available since 2005 when Yahoo launched a CLIR search interface option for the German and French sites [9]. Yahoo currently supports CLIR for 40 languages. On May 2007, Google launched its “Translated Search” in its Google Language Tools (http://www.google.com/language_tools) in addition to other language support services and tools. The launch of the cross-language search by Yahoo and Google was a breakthrough event because it signified the transition of CLIR from research to its real application. It’s probably the first time that CLIR and machine translation (MT) have been integrated to provide a real application on the Internet. We believe it is time to make CLIA services part of the development of digital libraries.

2. INTEGRATIVE CLIA FRAMEWORK

We are in the position to develop and evaluate an integrative CLIA (iCLIA) framework for digital libraries. The framework will extend the Lexical Knowledge Base (LKB) Model for English-Chinese Cross-Language Information Retrieval [1] to facilitate fast development of cross-language information access (CLIA) services for digital libraries. It will also include procedures and design principles for sustaining the CLIA services.

This framework will:

1. Integrate translational knowledge from multiple sources.

Translational knowledge is crucial for MT systems and query translation processes. Incomplete or faulty translational knowledge is the major cause of the out-of-vocabulary problem. In the iCLIA framework, translational knowledge should be obtained from multiple sources, such as commercial or open-source translation dictionaries, the DL collection, the Web, and the DL users. Efforts will be made to integrate translational knowledge from different resources to develop a lexical knowledge base (LKB) capable of translating queries and to augment the dictionary of MT systems for translating documents. Figure 1 is an illustration of the possible translational knowledge collection strategy: the translational knowledge will be collected from multiple sources including the DL collection itself through automatic information extraction, the possible commercial translational resources, the Web, other transnational resources such as open-source machine translation systems, and the users. Each source will be evaluated and assigned a label to specify its quality. The translational knowledge will be identified, organized, and stored in a lexical knowledge base for use during the retrieval stage.

Figure 1: the LKB Construction Process

Figure 2: The Structure of the LKB

The simplified structure of the LKB is showed in Figure 2. Translational knowledge can be stored in three relational tables.

---

Important terms (words or phrases) in the metadata or full text of the digital objects will be first extracted and stored in “Lexicon in A”, where “A” denotes the original language of the digital collection. Then the translational knowledge collection process as depicted in Figure 1 will be applied to generate “Lexicon in B” and the “A-B Mapping” tables. Here “B” denotes the query language of the users. The three tables will store important information such as the category of the terms and any associations among the terms. The translation probability will be determined by a formula considering the quality of various sources that provide the translation.

(2) Integrate computational power and human intelligence.

CLIA users are usually absent in CLIA literature. Our iCLIA framework emphasizes the interaction with DL users in order to improve the CLIA service for digital libraries. The CLIA service will be introduced to the users so that the whole user community of the digital library can participate in the development of the service. Then the CLIA process will include users who provide feedback on the translation and retrieval results. As illustrated in Figure 3, the CLIA process will be iterative and interactive: DL users will be able to ask queries in their languages, provide feedback on the translations returned by the system if they desire, and examine search results and their translation. The feedback from the users will be input into the system to update the LKB.

(3) Integrate various translational disambiguation approaches for best performance.

Translational disambiguation has been extensively explored in CLIR literature. Algorithms have been developed for CLIR experiments at IR evaluation forums such as TREC, CLEF, and NTCIR. However, agreement on a single, best approach has yet to be reached. The IR approach of combining results from multiple solutions may be interesting to explore for CLIA for DLs. Also, a translational disambiguation strategy may need to consider the language pairs to be processed.

(4) Integrate different measures for effectiveness and efficiency.

The CLIA services for digital libraries cannot copy IR performance measures such as precision and recall. More factors and criteria need to be considered in order to develop a dynamic, practical, and sustainable CLIA service. Performance and efficiency measures in system-oriented and user-oriented IR evaluation, library and digital library service evaluation, and usability consideration should all be considered in order to develop a comprehensive method for implementing CLIA services.

3. STRATEGIES

Collaboration is crucial in order to develop the iCLIA framework. The following strategies should be considered to develop the CLIA services based on the iCLIA framework for digital libraries with limited funding:

- Collaborate with Digital library (DL) developers to work on real digital libraries. Such collaboration facilitates the exploration of solutions that are appropriate for the specific digital objects while seeking funding to support CLIA as a value-added service. As digital objects are more organized than Web pages crawled by search engines, it is possible that better performance of machine translation could be achieved through the construction of the LKB for machine translation software.
- Collaborate with researchers and DL developers in other countries to increase the languages with which users can access documents. Many digital libraries manage precious digital assets that are often attractive to people on the other side of the earth. Collaboration with colleagues in other countries would make information resources available to a much larger population.
- Collaborate with the users. In the current digital age, even monolingual digital libraries are accessed by people using differing languages [8]. Social computing has been widely used on the Internet, and it can play a big role in involving users in the multilingual information access services: users may even volunteer to translate digital objects into another language. They may also help correct errors produced by machine translation systems. They will be more likely to donate money to help the DL to offer the new service if they know the significance of the service and the information needs from the other side of the earth.

Taking a step-by-step approach, the DL system can first be provided with a cross-language interface, then CLIA to the metadata, and then the full texts of the digital collections.

4. SUMMARY AND CONCLUSION

Information systems such as digital libraries would better serve their users if language support services were integrated as part of the systems. We propose to explore an integrative CLIA
framework for digital libraries. CLIA can be a value-added service to many DLs if we explore the methods, the user needs, and the evaluation of such services.

Our future research will focus on developing and testing the iCLIA framework to understand the application of CLIA for DLs. Within the framework, we aim to understand more about the needs and information behavior of bilingual users because bilingual users have been identified as the most probable users for CLIR systems. Also, we will collaborate with existing digital libraries to investigate effective and efficient solutions to providing multilingual information access for the digital library users.

5. REFERENCES
Ubiquity: Designing a Multilingual Natural Language Interface

Michael Yoshitaka Erlewine
Mozilla Labs
650 Castro Street
Mountain View, CA 94041
mitcho@mitcho.com

ABSTRACT
This paper describes the design and implementation of Ubiquity, a multilingual textual interface for the Firefox browser developed at Mozilla Labs. The Ubiquity interface facilitates rapid information retrieval and task execution in the browser, leveraging existing open web APIs. The importance of offering equivalent user experiences for speakers of different languages is reflected in the design of Ubiquity’s new natural language parser, described here. This paper also aims to advocate the further development of equipotent multilingual interfaces for information access.

Categories and Subject Descriptors
H.1.2 [Models and Principles]: User/Machine Systems—human information processing; H.5.2 [Information Interfaces and Presentation]: User Interfaces—natural language; I.2.7 [Artificial Intelligence]: Natural Language Processing—language parsing and understanding

General Terms
Design, Experimentation, Human factors, Languages

1. INTRODUCTION
Language continues to be one of the greatest barriers to open information access on the internet. The participation of ever more diverse linguistic communities on the web has not only created great linguistic divides in web content, but has also naturally resulted in a multitude of disparate tools created within each community, leaving such projects less able to benefit from each others’ innovations. While much effort and increased attention have been devoted to the development of multilingual corpora and resources, less attention has been given to guaranteeing that users with different linguistic backgrounds can use the same quality tools to access that information. As part of Mozilla’s goal to make the internet experience better for all users [8], Ubiquity aims to bring a new form of interactivity into the browser which treats user input in different languages equally. Ubiquity offers a platform for rapid information access, with no languages treated as second-class citizens.

The desire of users to access the internet using an interface in the language most natural to them is reflected in Mozilla’s latest Firefox browser release which shipped in over 70 languages, each localized by a team of volunteers. The goal of fulfilling this desire is particularly pertinent—and challenging—in the case of a natural language interface.

Ubiquity was born of the Humanized Enso product (http://www.humanized.com/enseo/), but is now an open-source community project, with dozens of contributors and active testers. It is available for download at http://ubiquity.mozilla.com and can be installed on the Firefox browser. Similar popular text-based command interfaces which are overlaid on GUI include Quicksilver (http://www.blacktree.com) and GNOME Do (http://do.davesbd.com/), but neither of them attempts a natural language syntax, nor do they support localization of their parser and keywords.

2. TOWARDS A NATURAL INTERFACE
2.1 Features of a Natural Syntax
The lead of Ubiquity development Aza Raskin argues in his 2008 ACM interactions paper that text-based interfaces can be more humane than overextended graphical interfaces [10]. Graphical interfaces are easy to learn and apply for concrete tasks but do not scale well with additional functionality and lack the precision required to communicate abstract instruction. While numerous text-based computer interfaces exist, they have been deemed too difficult for lay users. Raskin argues that textual interaction does not entail these difficulties per se; rather, they are products of their oft-times stilted grammars. In reconsidering the text-based interface, ease and familiarity built into the interface are key. A subset of natural language is thus a clear winner.

Many programming and scripting languages—themselves interfaces to instruct the computer—make use of keywords inspired by natural languages (most often English). Many simple expressions neatly mirror a natural language (1a) but more complex instructions will quickly deviate (1b).

\[(1) \ a. \ \textbf{print} \ "\text{Hello World}" \quad (\text{Python})\]
\[\quad b. \ \textbf{print} \ \textbf{map}(\lambda x: x*2, [1,2,3])\]

\[1\text{The term “humane” is used in this paper to describe human-computer interfaces which are “responsive to human needs and considerate of human frailties” [12] (see also [11]).}\]
One valiant effort to facilitate near-natural language instruction has been AppleScript, which enables complex English-like syntax (as in 2) and originally was planned to support similar Japanese and French “dialects.”

(2) print pages 1 thru 5 of document 2 (AppleScript)

As a full-featured scripting language, however, more complex expressions push beyond the natural language metaphor and introduce their own idiosyncrasies. Bill Cook, one of the original developers of AppleScript, notes “in hindsight, it is not clear whether it is easier for novice users to work with a scripting language that resembles natural language, with all its special cases and idiosyncrasies” [5]. Raskin notes that this is precisely what must be addressed in designing a humane text-based interface: “if commands were memorable, and their syntax forgiving, perhaps we wouldn’t be so scared to reconsider these interface paradigms” [10].

In designing an internationalizable natural language interface, we can conclude that it is not enough to use natural language keywords and mimic its syntax. The grammar must never conflict with a user’s natural intuitions about their own language’s syntax—a goal I call natural syntax. While a user can’t expect such an interface to understand every natural language command, a good rule of thumb is that multiple natural alternatives for a given intent are interpreted in the same way. For example, consider the examples (3) in Japanese, a language with scrambling:

(3) a. 太郎に ボールを 投げる
   Taro-ni ball-o nagero
   Taro-DAT ball-ACC throw-IMPER

b. ボールを 太郎に 投げる
   ball-o Taro-ni nagero
   ball-ACC Taro-DAT throw-IMPER

Both sentences are valid expressions for the command “throw a ball to Taro.” An interface with a natural syntax must understand either both of these inputs or, if for example the interface does not understand the verb nagero, neither of them. To understand one but not the other goes against the tenet of natural syntax.

2.2 Commands in Ubiquity

Ubiquity actions are requests for actions or information, corresponding functionally to the formal clause type of “imperative” [9], although they may manifest in forms traditionally characterized as “imperative,” “infinitive,” or “subjunctive,” depending on the language [7]. No vocative is entered as the addressee is always the computer, nor do we handle negation, leaving Ubiquity input simply to be composed of a single verb and its arguments (if any). Some example English Ubiquity actions include:

(4) a. translate hello to Spanish—previews the text “hola.” On execution, inserts the text “hola” in the active text field.

Figure 1: Schematic diagram of user interaction with Ubiquity.

b. email hello to John—on execution, composes a new email to contact John with message body “hello.”

c. map Tokyo—previews a map of Tokyo using the Google Maps API. The image can then be inserted into the page.

Verbs are written in JavaScript. Each verb may specify a preview() method which displays some information to the user or gives a preview of the action to be executed and an execute() method which carries out the intended action.

In order to avoid ambiguity, a list of possible parses is presented to the user for confirmation before execution. Suggestions give a visual indication of the parsing. A scoring mechanism is used to bring more likely candidates to the top, taking user input and browser use habits into consideration.

3. ADDRESSING THE NEEDS OF MULTILINGUAL ACCESS

With the requirements and goals of the project as laid out in section 2, certain architectural choices were made in designing the parser in order to support multiple languages equipotentially. In this section I will review the unique features of our parser and platform which enable equal information access and rapid localization.

3.1 Identifying Arguments by Semantic Role

Ubiquity commands’ ease of creation is a great strength for the platform, with many contributors around the world creating and sharing their own verbs as well as writing new verbs for personal use. In order to let users of different languages benefit equally from the platform, however, there is a
need to internationalize the verbs themselves. Verbs include some strings which must be translated, such as the verb’s name, but they also include a specification of the type of arguments it accepts, known as the syntactic frame of the verb. For example, in English an email verb may take a direct object and a recipient introduced by the preposition “to,” while a translate verb may take an arbitrary direct object, a goal language marked by “to,” and a source language marked by “from.”

In order to facilitate this localization, we chose to let verbs specify their syntactic frames using abstract semantic roles such as object, goal, instrument, position, etc. which are morphosyntactically coded in most languages. For example, suppose an English-speaking contributor wrote a verb called move, whose action was to move an object from one location to another. Its syntactic frame could be specified as follows, where physical_object and location are noun types which specify a class of arguments and their associated semantic forms.

{ object: physical_object, source: location, goal: location }

The command author could then use this command in English, entering input such as (5). The parser recognizes the English prepositions “to” and “from” as corresponding to the goal and source roles (underlined below), and recognizes the unmarked argument as an object.

(5) move truck from Paris to Beijing

(6) トラックをパリから北京へ搬

However, given a set of localized noun types, the exact same command code could be used with the Japanese parser by entering the input (6). Here, the parser recognizes that the postpositions “を,” “へ,” and “から” mark object, goal, and source arguments, respectively. The only manual localization required for the move command, then, is the translation of the verb name “move” itself. As shown by this example, the specification of arguments using abstract semantic roles supports the rapid and, indeed, semi-automatic localization of commands, ensuring that users of all languages benefit from individual improvements to the Ubiquity platform’s functionality.

### 3.2 Argument-first Suggestions

In parsing Ubiquity input, a key task is the identification of the verb, if any. In many languages the verb naturally comes at the beginning of the sentence (see English examples in 4). In this case, as the verb can be identified early in the user input, we can then annotate the candidate parses with information on the missing arguments to guide the user in entering the rest of their input (see figure 2). However, not all languages enter the verb first in commands. Some languages are strictly verb-final (e.g. Japanese), while in some other languages (e.g. German, Dutch, Modern Greek) it is equally valid to express commands using the imperative or subjunctive verb form at the beginning of the sentence, or using the infinitive at the end of the sentence.

Rather than being discouraged by this conundrum, thought was given to how we can leverage the unique qualities of verb-final (or argument-first) input to make a more humane and supportive interface. As different verbs in our lexicon specify different syntactic frames, by parsing arguments and identifying semantic roles in the input before the verb is known, we can then suggest verbs to the user which match that particular argument structure (see examples in table 1 of some such suggestions). This smart argument-first suggestion aids in command discoverability by suggesting verbs for a given target which the user may not have known.

<table>
<thead>
<tr>
<th>Sample argument parses</th>
<th>suggested verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>{ object:..., goal:... }</td>
<td>email, send</td>
</tr>
<tr>
<td>{ object:..., instrument:... }</td>
<td>search, look up</td>
</tr>
<tr>
<td>{ object:..., source:..., goal:... }</td>
<td>move, translate</td>
</tr>
</tbody>
</table>
exists. This approach crucially takes advantage of the argument-first input and offers unique value and increased usability to users with verb-final languages.

Note also that the suggestion of verbs based on argument-only input may also be useful for regularly verb-initial languages such as English. Studies of general interactive systems concur that noun-verb (or object-action) paradigms result in error reduction, increased speed, and better reversibility during input [11]. For these reasons, argument-first suggestions are supported in Ubiquity for all languages equally.

3.3 Minimal Language Descriptions

The Ubiquity parser attempts to make as much of its parser algorithm universal as is practical, taking a page from the Principles and Parameters framework in generative linguistics. A single universal parser was designed, with settings for different languages built on top of that base [1]. The settings for each language are written in JavaScript and range from ten to thirty lines of code. Various hooks exist in the code for language-specific processing when necessary, but the majority of the language settings are simply lists of special lexical items such as the prepositions or postpositions in a language. In this way, for the limited range of data which constitute Ubiquity input, the very difficult problem of writing a language-specific parser is reduced to little more than some native speaker consultation and string translation.

4. EVALUATION METRICS

As an open-source community project, the success of Ubiquity must be evaluated in terms of developer involvement as well as user adoption. The design choices laid out here are intended to lower the barrier of contributing to Ubiquity’s localization, and the level of localization engagement is a direct reflection of the facility or difficulty of contributing to individual parser language setting files and localizations. An initial survey of this metric can be interpreted as optimistic, with language settings having been written for ten languages and Ubiquity’s built-in commands having been completely localized into three of those languages as of this writing, even before widespread public release of the new parser.

Approximate user adoption rates can be calculated based on download and update counts, though this does not currently directly reflect active usage of Ubiquity nor give us much insight into its different use cases. Current user adoption is expected to be limited by the fact that previous versions of Ubiquity were almost exclusively suited for English use, and we expect a slow uptick in usage and interest by users in other languages which we are beginning to support. Future collaboration with the Mozilla Labs’ Test Pilot project [4] to collect anonymous user behavior data in Ubiquity is also being planned [3]. This data will help yield more accurate usage statistics, including usage breakdowns by language, as well as yield valuable information on parser accuracy and user interaction patterns.

5 It is worth noting that this architectural choice also complemented the object-oriented architecture and Don’t Repeat Yourself design goals of the project.

5. CONCLUSIONS

Further globalization of the web without serious consideration of multilingual information access could spur the further fragmentation of information and ideas. Equal access to information will require more than just cross-language search and retrieval systems, but also universal interfaces which are designed for rapid localization and treat all languages equally.

In this paper I outlined some of the design features of Ubiquity’s interface and natural language parser which bring the system closer to this goal. Formal approaches to the study of language were applied in order to design a system which can be extended to a wide range of languages. As of this writing, settings for ten languages have been written for Ubiquity, while the community process of setting technical standards for verb and noun type localization is in progress.

Ubiquity is quickly becoming a compelling text-based interface for both advanced and casual users. The forgiving “natural syntax” philosophy and the smart suggestion of verbs and arguments to the user help make Ubiquity a humane interface which cooperates with users rather than confounds them. These qualities make Ubiquity a natural choice of interface platform for multilingual and cross-language information access applications.

6. ACKNOWLEDGMENTS

Thank you to comments from Aza Raskin and Jonathan DiCarlo at Mozilla and audiences at BarCamp Tokyo; Chuo, Waseda, and Keio Universities; Tokyo 2.0; Tokyo Institute of Technology; as well as comments on related material on my blog (http://mitcho.com/blog/).

7. REFERENCES


July 23, 2009 48 Gey, Kando, Karlgren, Chairs
Romanization – An Untapped Resource for Out-of-Vocabulary Machine Translation for CLIR

Fredric Gey
University of California, Berkeley
UC Data Archive & Technical Assistance
Berkeley, CA 94720-5100
510-643-1298
gey@berkeley.edu

ABSTRACT

In Cross-Language Information Retrieval (CLIR), the most continuing problem in query translation is the occurrence of out-of-vocabulary (OOV) terms which are not found in the resources available for machine translation (MT), e.g. dictionaries, etc. This usually occurs when new named entities appear in news or other articles which have not been entered into the resource. Often these named entities have been phonetically rendered into the target language, usually from English. Phonetic back-transliteration can be achieved in a number of ways. One of these, which has been under-utilized for MT is Romanization, or rule-based transliteration of foreign typescript into the Latin alphabet. We argue that Romanization, coupled with approximate string matching, can become a new resource for approaching the OOV problem within parallel corpora or enter into formal dictionaries. In addition, a plethora of name variants also confuse the issue of named entity recognition. Steinberger and Pouliquen (2007) discuss these issues in detail when dealing with multilingual news summarization. For non-Latin scripts, this becomes particularly problematic because the user of western scripted languages (such as in USA, Europe) cannot guess phonetically what the name might be in his/her native language, even if the word or phrase was borrowed from English in the first place. In many cases, borrowed words enter the language as a phonetic rendering, or transliteration or the original language word. For example, the Japanese word コンピュータ (computer), Knight and Graehl (1997) jump-started transliteration research, particularly for Japanese-English by developing a finite state machine for phonetic recognition between the two languages. The phonetic transliteration of the above Japanese is ‘kompuyuta’.

There is, however, an alternative to phonetic transliteration, and that is Romanization, or rule-based rendering of a foreign script into the Latin alphabet. Romanization has been around for a long time. For Japanese, the Hepburn Romanization system was first presented in 1887. The Hepburn Romanization for the Japanese ‘computer’ above is ‘kompuyuta’. The Hepburn system is widely known that a PERL module for Hepburn is available from the CPAN archive.

In addition to Hepburn, there has been a long practice by the USA Library of Congress to Romanize foreign scripts when cataloging the titles of books written in foreign languages. Figure 1 presents a list of about 55 languages for which the Library of Congress has published Romanization tables. Note that major Indian subcontinent languages of Bengali, Gujarati, Hindi, Marathi, Punjabi, Tamil, Telugu and Urdu are included. For example, the Cyrillic Клинтон or the Greek Κλίντον can easily be Romanized to Clinton. For Russian and Greek, the transformation is usually reversible. For the major Indian language, Hindi, it is easily possible to find the translation for Clinton, but for the south Indian language of Tamil, translations are less easily found. Yet Tamil is a rather regular phonetic language and foreign names are often transliterated when news stories are written in Tamil (although one reviewer has remarked that Tamil has phonetic ambiguities not found in other Indian languages). Figure 2 is a translated news story in Tamil, when the main names (Presidents Clinton and Yeltsin) are Romanized.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. SIGIR Workshop on Information Access in a Multilingual World July 23, 2009 Boston, Massachusetts USA.
2. TRANSLITERATION/ROMANIZATION

In the sweep of methods for recognition of out-of-vocabulary terms between languages and for automatic phonetic recognition of borrowed terms, Romanization has become a much-neglected stepchild. However phonetic transliteration (and back-transliteration from the target language to the source language) requires large training sets for machine learning to take place. For less-commonly taught languages, such as, for example, some Indian subcontinent languages, such training sets may not be available. Romanization, on the other hand, requires that rules for alphabet mapping be already in place, developed by experts in both target and source languages. However, once the target language word has been rendered into its Latin alphabet equivalent, we still have the problem of matching it to its translation in the source language. So we ask: Is there a place for Romanization in CLIR? And how can it be exploited? The key is the examination of approximate string matching methods to find the correspondences between words of the target and source languages.

3. APPROXIMATE STRING MATCHING

Once one has Romanized a section of non-English text containing OOV, the task remains to find its English word equivalents. The natural way to do this is using approximate string matching techniques. The most well-known technique is edit distance, the number of insertions, deletions and interchanges necessary to transform one string to its matching string. For example, the edit distance between computer and kompyuta (コンピュータ) is 5. Easier to comprehend is between English and German, where the Edit distance between fish (E) and fisch (DE) is 1. However, the edit distance between fish(E) and frisch (DE) is 2, whereas between the correct translations fresh (E) and frisch (DE) is also 2. Thus Martin Braschler of the University of Zurich has remarked, “Edit distance is a terrible cross-lingual matching method.” Approximate string matching has a lengthy history for both fast file search techniques as well as finding matches of minor word translation variants across languages. Q-grams, as proposed by Ukkonen (1992) counts the number of substrings of size ‘q’ in common between the strings being matched. A variant of q-grams are targeted s-grams where q is of size 2 and skips are allowed to omit letters from the match. Pirkkola and others (2003) used this technique for cross-language search between Finnish, Swedish and German. Using s-gram skips solves the fish – fisch differential above. An alternative approach, which has been around for some time, is the Phonix method of Gadd (1998) which applies a series of transformations to letters (for example, c → k, in many cases, e.g. Clinton → Klinton) and shrinks out the vowels, (Clinton → Klint). If we apply this transformation to the English Japanese above, we have computer → kmpt and compyuta → kmpt. The original version of Phonix only kept the leading four resulting characters, and would result in an exact match. Zobel and Dart (1995) did an extensive examination of approximate matching methods for digital libraries and their second paper (1996) proposed an improved Phonix method they titled Phonix-plus which did not truncate to 4 characters, but instead rewarded matches at the beginning. They combined this with edit distance for the Zobel-Dart matching algorithm.
4. SUMMARY AND POSITION

The current fashion for utilizing statistical machine learning as the solution to all problems in machine translation has led to the neglect of rule-based methods which, this paper argues, are both well-developed and could complement statistical approaches. Romanization would work especially well for non-Latin scripted languages for which training corpora are limited. The approach has two steps: 1) Romanization of the script using well-documented methods, followed by 2) Approximate string matching between Romanized words in the target language and possible translation candidates in the source language.

5. ACKNOWLEDGMENTS

Much of this work was originally done while the author was a visiting researcher at the National Institute of Informatics (NII) in Tokyo in the summer of 2007 supported by a grant from NII.

6. REFERENCES


Figure 2: News story in the Tamil language of Clinton-Yeltsin Meeting, showing name Romanization (phonetic transliteration according to software from the University of Cologne, Germany)
Urdu is not Hindi for Information Access

Kashif Riaz
University of Minnesota
4-192 EE/CS Building
200 Union Street SE
Minneapolis, MN 55455
riaz@cs.umn.edu

ABSTRACT
Urdu and Hindi have a complex relationship. Urdu is written in the Arabic script, and Hindi is written in Devanagri script. The relationship between these two languages is deep-rooted based on geo-political and historical events. Urdu and Hindi are considered very similar languages, and sometimes linguists refer to them as Hindi-Urdu. Urdu is considered a “scarce resource” language but Hindi has a vibrant toolset to do research in Information Retrieval.

In this position paper I contend that language resources and enabling technologies for Information Access cannot be used interchangeably between these two languages. More specifically, Hindi cannot be used as bridge language to do research in Information Retrieval in Urdu. I argue this assertion using deep analysis of the language through socio-linguistics and quantitative analysis using Zip’s Law. The contrast and comparison are done using script and vocabulary differences between the two languages.

Categories and Subject Descriptors
H.3.1 [Content Analysis and Indexing]: Linguistic Processing, Dictionaries, Indexing methods.

General Terms
Algorithms, Standardization, Languages

Keywords
Urdu, Hindi, Script, Orthography, Language Resources

1. INTRODUCTION
Urdu and Hindi share a complex relationship. Together they boast one of the largest populace who understands them and call either one of them as their national language. They are the languages of South Asia—Urdu is the national language of Pakistan, and Hindi is the official language of India. India does not have a national language because of the number of regional language-sensitive provinces. Urdu is one of the official languages of India. Urdu is written in the Arabic script, and Hindi is written in Devanagri script. While doing research on Urdu named entity recognition, I wanted to use some Hindi language resources like gazetteers and online dictionaries since they are not available for Urdu. I realized in early research stages that Hindi and Urdu were behaving as separate languages. Moreover, I needed to learn Devanagri script to proceed further. In this position paper I argue that Hindi and Urdu, although grammatically very similar, cannot be treated as the same language for doing research in computational linguistics and information retrieval—at least with the current set of tools available for both languages. Hindi has quite a vibrant set of enabling technologies for Information Access whereas research in Urdu is still in its infancy. Some of the examples of these enabling technologies which are available for Hindi are: online dictionaries, Wordnet, stemmers, stop word lists [14], gazetteers for named entity recognition systems [14], part of speech taggers [14], baseline for evaluation like FIRE [15] to name a few. Some of the mentioned tools for Urdu are available through CRULP [4] and [1, 3, 9, 11, 12] in varying development stages. Section 2 gives a brief overview of Urdu. Section 3 analyzes the differences between Urdu and Hindi and motivates why a position needs to be taken for doing Information Retrieval. Section 4 gives three examples of quantitative analysis between Hindi and Urdu to reinforce the position. Section 5 concludes the position paper with remarks.

2. Urdu
This section briefly introduces some right-to-left languages and a few characteristics of Urdu. Urdu is the national language of Pakistan and one of the major languages of India. It is estimated that there are about 300 million speakers of Urdu. Most Urdu speakers live in Pakistan, India, the UAE, the UK, and the USA. Urdu is considered the lingua franca of business in Pakistan and amongst the South Asian community in the UK [3]. Urdu has a property of accepting lexical features and vocabulary from other languages, most notably, English.

Recently there has been quite a bit of interest in right-to-left language processing in the IR community—specifically in the intelligence community. Most of the interest has been focused toward Arabic, Urdu, Persian (Farsi), and Dari. Arabic is a Semitic language, and the other languages belong to the Proto Indo Iranian branch of languages. Arabic and these other languages only share script and some vocabulary. Therefore, the language-specific task done for Arabic is not applicable to these languages. For example, stemming algorithms generated for Arabic will not work for a language like Urdu.

Urdu, among all above languages mentioned, has unique case in that it shares its grammar with Hindi. The difference is some vocabulary and writing style. Hindi is written in Devanagri script. Because of the grammatical similarity, Hindi and Urdu are considered one language with two different scripts by many linguists. I will argue in later sections that there is a growing number of dissenters among South Asian language researchers. Urdu is quite a complex language because Urdu’s grammar and morphology is a combination of many languages: Sanskrit, Arabic, Farsi, English, and Turkish to name a few. This aspect of Urdu becomes quite challenging while building Information Access systems [9]. Because of its rich morphology and word-borrowing characteristics, Urdu is widely considered the language
of the poets. From the Mughal courts of the 1700s to freedom writings of the 1900s, it was a prerequisite to learn Urdu in order to be considered a reputable poet or intellectual.

3. Analysis of Hindi and Urdu

3.1 Background

While doing research on Urdu Information Access, a claim can be made that Urdu and Hindi are the same language in two different scripts. According to this theory, any computational model or algorithm that works for Hindi should also work for Urdu. Some examples are baseline for evaluation methodologies that are designed for Hindi [15], gazetteers for name entity recognition, and Wordnet for Information Access. The following section describes in detail that the one language two scripts theory for Urdu and Hindi is invalid in all circumstances and specifically for computational processing.

Although a lot of research has been done about the origins of Urdu and Hindi, no research study exists that compares and contrast Urdu and Hindi in a scholarly fashion [13]. For quite some time, Urdu and Hindi were treated as the same language and, indeed, they are very similar. For geo-political reasons, the languages can be classified as two languages. The geo-political reasons are of no concern to this study, but they play an important role in why these two languages are currently diverging. Only linguistic and pragmatic reasons should be considered while studying the nature of Hindi and Urdu and their impact on computational processing. I will also exclude the socio-linguistic aspects in terms of religion (i.e., Hindi for Hindus and Urdu for Muslims in India) because this leads the discussion elsewhere.

With my linguistic training, I used to think that Urdu and Hindi were the same language and differed in vocabulary only. Hindi vocabulary emerges from Sanskrit, and Urdu vocabulary borrows from many other languages, but the majority of the borrowing comes from Persian and Arabic. Some rudimentary experiments for computationally recognizing names show that Hindi and Urdu behave as two different languages. For example, lexical cues of recognition of locations are different. For example, Dar-al-Khilafah (Urdu) and Rajdhani (Hindi) are both used for the capitol of a city or a country. Therefore, more research is warranted to understand the relationship between these two languages to understand if the computational models based on one language can be used in some capacity for the other language. For this study, my reasoning is based on the analysis of some of the revered scholars of the Urdu and Hindi languages, like Ralph Russell, David Matthews, Robert King, and Intizar Hussain.

3.2 Hindustani is not the Predecessor of Hindi and Urdu

Some researchers claim that Urdu and Hindi emerged as the offspring of the earlier language called Hindustani. Although there is some reference to a language called Hindustani in the later 1800s and early 20th century, a deeper analysis shows that no such language existed, and if it did, it was Urdu. Urdu was chosen as the official language of India by the British—that changed only after the Mutiny of 1857 [6]. The use of the word “Hindustani” emerged from the leaders of the Congress Party, primarily Gandhi, and then Nehru, who wanted a common language for the united India. From the early 1900s, Gandhi relentlessly pursued the theory that the people of northern India spoke neither Persian-based Urdu nor Sanskrit-based Hindi; instead they spoke Hindustani. The reason for this argument by the party leaders was that there needed to be one language for the united India. Gandhi, and later Nehru, never talked about which script would be chosen because of the intense emotions attached to the issue [6]. This political desire did not culminate as planned because most of the followers of the Congress Party wanted the Devanagri script to be the official script for Hindustani, an idea which was rejected by the mostly Urdu-speaking populace in northern India and Hyderabad. The use of the term Hindustani after the partition of India and Pakistan was used to describe the political tension around language choice during the British Raj.

3.3 Divergent Trend

Hindi and Urdu have two very clear differences: script and vocabulary. I will discuss the vocabulary differences in this section. Recent research in Urdu and Hindi studies is consistently arguing that Hindi and Urdu are two different languages and that they continue to diverge as time goes on. The most notable example of continual diverging is the ultra sanskritizing of Hindi so much so that an Urdu speaker does not understand a Hindi news broadcast. This does not mean that Urdu and Hindi speakers don’t understand each other; they do at an everyday level. Also, both Hindi and Urdu speakers who live together in Uttar Pradesh, Andhra Pradesh, and other large cities understand both languages effortlessly. But a person from an outside region who travels to these areas and knows only one language does not understand the other language [6][8][13]. Matthews [8] and Martynyuk [7] make a similar comparison with Russian, Ukrainian, and Serbo-Croatian languages in terms of their similarities in common vocabulary but still being different languages. King [6] claims that the relationship between Urdu and Hindi is much more complex than Cyrillic orthography-based Serbian and romanized Croatian collectively known as Serbo-Croatian. Given the discussion above, common vocabulary claim cannot be used to claim that two languages are same.

The relationship between Hindi and Urdu is very complex. While analyzing the differences at a high level, they can be treated as the same language and play a pivotal role in establishing a link between South Asian communities around the world. A glowing example of this phenomenon is the Indian cinema where the line between Hindi and Urdu gets diluted. Although Hindi movies are popular in Urdu-speaking Pakistan and Pakistani TV shows are popular in India, there is steady and noticeable shift in Indian movies towards sanskritizing of Indian cinema. Remarkably, the Indian movies produced from the 1950s to the 1980s are undoubtedly Urdu (e.g. Pakeezah made in 1972 and sanskritized Swades made in 2004). At a detailed level, Urdu and Hindi are separate languages and deserve to be studied and treated as separate languages. This is most apparent in the official documents produced by the Indian government in Hindi and news broadcasts [8]. Noticing the growing trend of the usage of Sanskrit words in Hindi, researchers of both Urdu and Hindi have started to describe them as separate languages. The commonality of the two languages is described by Matthews [8] as an unfortunate oversimplification of two vibrant languages. Russell [13] in his critique of Christopher King’s book, “One Language Two Scripts”, cites a number of examples where he shows that Hindi and Urdu are similar but different languages and sometimes

1 Sanskritization is defined by anthropologist as spread of Vedantic and Brahmanical culture.
the vocabulary, usage, and pronunciation can make a huge impact on understanding of the language. Russell compared language teaching books of Hindi and noted that a number of the words can easily be treated as Urdu like akela (alone) and akelapan (loneliness) but soon the difference start to appear (e.g., adhik (lots) in contrast to zyada (lots) in Urdu, akash (sky), asman(sky) in Urdu).

A few examples like these should not be used to make a statement about two different languages. The translation of “How far is your house from here?” will be understood both by Hindi and Urdu speakers, but the divergence trend between these two languages continues with time (e.g., bicycle in Hindi referred to as do chakr ghamia which was heard by the author in Hindi news telecast whereas in Urdu it remains to be called cycle). The official usage is mostly present in the speech transcriptions of the Indian Parliament and the official government documents. The following example is borrowed from Russell to explain the growing divergence. Consider the sentence in English “The eighteenth century was the period of the social, economic and political decline”. The Urdu translation of the sentence is “Atharvin sadi samajik, igtsadi aur siyasi zaval ka daur tha” while the Hindi equivalent is “Atharvin sadi samajik, arthik aur rajnitik girav ki sadi thi”. Moreover, in Hindi “sadi” could be replaced by “satabdi” and “aur” with “tatha”. Russell points out that this example alone shows that Urdu speakers cannot understand the meaning of the Hindi equivalent and vice versa. Therefore, these two languages should not be treated as the same language in all circumstances.

3.3.1 Highbrow, Middlebrow, and Lowbrow

Besides script, the most notable differences between Hindi and Urdu are found in the formalized vocabulary, grammar, and writing style. King [8] quoting Ashok Kelkar, a proponent of Hindi, describes those differences in detail as an excellent example of a social linguistics situation. Hindi and Urdu have a full range of styles. He categorized those styles as stated below:

- **Formalized highbrow** is used in academia, religious sermons, official texts, and poetry. Most language engineering resources and enabling technologies for system development are based on this style. Highbrow Hindi draws its base from Sanskrit and highbrow Urdu, throughout time, has been based on Persian and Arabic words.

- **Formalized middlebrow** is used in songs, movies, pamphlets, popular printed literature, and mass propaganda.

- **Casual middlebrow** is most widely used for daily conversations among the educated upper and middle class who are regionally based like in northern India and Hyderabad. It is used for private communication like phone conversations and letter writing. It is used by newspapers so they can be read by a large audience. This style is most receptive to borrowed words, most of them from English.

- **Casual lowbrow** is associated with what Kelkar calls the “lower class” and uneducated people. He calls it bazaar Hindustani. This is definitely a substandard form of style. This style is found in the slums of urban centers of large Indian cities.

The polarization of Urdu and Hindi reaches its maximum at **formalized highbrow**. Hindi draws from Sanskrit for vocabulary and promotes Vedantic and Brahmansical culture. Urdu draws from Turkish and classical Persian literature and Islamic events as references.

King [6] mentions that standard Hindi (highbrow) and standard Urdu (highbrow) have diverged more since the partition of India and Pakistan in 1947. A careful analysis of King’s theory shows that it is certainly true that standard Hindi is getting more and more sanskrititized, but new Urdu literature is leaning towards formalized middlebrow. Sanskritized Hindi is increasingly used by the elite in India. This movement of sanskritizing Hindi in India is illustrated by King [6] while quoting Das Gupta and Gumperz. The illustration is done by analyzing the signboards; label a is the official text of the signboard, label b is the English translation and label c is the causal middlebrow in Hindi. I have added label d as the highbrow in Urdu and label e is the middlebrow in Urdu.

- **Signboard 1**
  a. dhumpan varjint hai
  b. smoking is prohibited
  c. cigarette pina mana hai
  d. tambakoo noshi mana hai
  e. cigarette pina mana hai

- **Signboard 2**
  a. Bina agya pravesh nishedh
  b. entrance prohibited without permission
  c. bina agya andar jana mana hai
  d. baghair poochey andar aana mana hai
  e. baghair poochey andar aana mana hai

Note that for signboard 2 middlebrow and highbrow is the same for Urdu as suggested earlier.

3.3.2 Cultural differences

Although languages don’t belong to a religious group, it is an undeniable fact that Urdu is the first language of the large Muslim population in India and is known to most Muslim Pakistanis as their national language. The same is true for the Hindu majority in India where most Hindus in North India prefer Hindi. There are of course, Hindu scholars of Urdu—Prem Chand and Gopi Chand Narang are notable examples.

The cultural preferences of speakers translate into respective languages. The date and year reference for Muslims for major events in South Asia is the Hijri calendar (reference to Prophet Muhammad’s migration from Makah to Medina). The year 2000 in C.E. is 1421 Hijri. This is evident in how different people reference the completion of the Taj Mahal for example. A Muslim cleric in Aligarh will refer to its completion in 1076 Hijri, but a secular Hindu will say the date is 1666 A.D. The epitaph inside the Taj Mahal refers to the Hijri date not the Gregorian calendar. A secular Hindu will say the date is 1666 A.D. The epitaph inside the Taj Mahal refers to the Hijri date not the Gregorian calendar.

The architectural terminology is quite different. A secular Hindu’s description of the Jama Masjid (Central Mosque) in Delhi will be quite different than a local Muslim. For example, a secular Hindu might not be able to explain alignment of mimbar (a platform for the prayer leader) in the mosque or know what it symbolizes. Religious symbols can also be confusing e.g., a Muslim in Pakistan will not know the importance of the river Ganga for religious Hindus. Mimbar and Ganga are used considerably with their implied meanings in Urdu and Hindi literature.
3.4 Script Differences

In this section I will explain few of the script differences between Hindi and Urdu in terms of phonemes, spoken units of a language, and graphemes, written units of a language. Hindi and Urdu have most of the phonological features of the languages of the subcontinent like retroflexion and voiceless and voiced, aspirated and unaspirated stops. The majority of the differences in Urdu and Hindi regarding the script are based on the vocabulary. Besides supporting the features of the languages of the sub-continent, the Urdu script supports the phonemes of Persian and Arabic. For example, in contrast to Hindi, Urdu has an unaspirated uvular stop /q/, labial fricative /h/, voiceless retroflex /ʃ/, velar fricative /x/, voiced dental fricative /l/, palato-alveolar voiced fricative /ʒ/, and voiced velar fricative /v/. These sounds are supported by the nastaliq and naksh styles of Urdu script. Hindi has a system of making these sounds native by changing the articulation at different levels for each foreign sound. Urdu script has distinct graphemic features for retroflexion and aspiration. Urdu uses diacritic marks for retroflexion, and aspiration in Urdu is shown by h whereas the Devanagri script of Hindi does not treat retroflexion and aspiration as distinctive features. There are a number of other examples, but the few examples above justify the difficulty when using Hindi resources for Urdu computational processing. One of the easier tasks for language engineering is a transliteration from one language to another by using a map from one symbol to another. The work of Jawaid and Ahmed [5] shows that there are many open issues when transliterating Hindi to Urdu or vice versa. The above differences show that Hindi stemmers cannot be used for Urdu stemmers [9]. The word fiqh (jurisprudence) in Hindi has labial fricative, uvular stop, and, aspiration. A Hindi morphological analyzer is probably going to fail while parsing the token fiqh.

4. Quantitative Analysis of Hindi and Urdu

In this section a few quantitative examples where Hindi resources cannot be used for Urdu Information Access are shown.

4.1 Named Entity Recognition

Named entity recognition (NER) is one of the important tasks in the field of Information Access. It constitutes automatically recognizing proper nouns like people names, location names, and organization names in unstructured text. There has been significant work done in English and European languages, but this task is not well-studied for the languages of South Asia.

By far the most comprehensive attempt made to study NER for South Asian and Southeast Asian languages was by the NER workshop at the International Joint Conference of Natural Language Processing in 2008 [14]. The workshop attempted to do named entity recognition in Hindi, Bengali, Telugu, Oriya, and Urdu. Among all these languages, Urdu is the only one that has Arabic script. There are fifteen papers in the final proceedings of the NER workshop at IJCNLP 2008. A number of those papers tried to address all South Asian languages in general, but resorted to Hindi where the most number of resources were available. There was not a single paper that focused on Urdu named entity recognition. In the papers that tried to address all languages, the computational model showed the lowest performance on Urdu. None of the researchers were able to use the online dictionaries and gazetteers that were available for Hindi for Urdu.

4.2 Zipf’s Law on Hindi and Urdu

One of the most interesting, and maybe the only, works available that compares Urdu and Hindi quantitatively is by Martynyuk [7]. In this study Martynyuk establishes his argument based on the comparative analysis of the most frequent words used in various European languages. He then extends his work to compare Hindi and Urdu based on Zipf’s Law. Zipf’s law is one of the most fundamental laws used by researchers when analyzing text in an automated way. On a high level, Zipf’s law states that given a corpus of words, the frequency of the words is inversely proportional to the rank of the word. Although not stated clearly, the hypothesis of the experiment is that if the most frequent words show similar rank order in corpora of different languages like Hindi and Urdu, then the theory of “same language two scripts” has a good chance of being accepted in the larger academic community. Hindi and Urdu news text tokens were categorized using manual lemmatization. The experiment was conducted on approximately 440,000 Romanized words of Urdu and Hindi. After calculating the frequencies of the most frequent words, it was found that the top three words between Urdu and Hindi were the same words (stop words). The rank order between the words starts to divert after rank three. The top 24 ranked words are the same between the two languages, but have a different rank order than each other. After rank 24, the words start to differ in rank order and their alignment with the corresponding word from the other language. For example the Hindi word for election, chunaav, is at rank 25; the Urdu equivalent for election, intikhabaat, is found at rank 45. It is important to note that the Hindi corpus did not contain the English word election, but in the Urdu corpus it was used almost as many times as intikhabaat (intikhabaat is used 672 times and election is used 642 times.) This next example may drive home the point that the rank order of the same meanings keep on drifting apart. The word for terrorist in Hindi, atankvadee, is at rank 42 in contrast to the equivalent word in Urdu, dehshatgard, which is at rank 182. This experiment clearly shows the Information Access features like term frequency, inverse document frequency, and stop word analysis for Hindi and Urdu will show different results because the difference in top ranked words [12].

4.3 Hindi Wordnet

This research is part of the larger research area doing concept searching in Urdu [10]. Wordnet is an important enabling technology for concept understanding and word sense disambiguation tasks. Hindi Wordnet [2] is an excellent source for Hindi language processing but cannot be used for Urdu. Most of the analysis of the words and the categorization of words in Hindi Wordnet is done by using highbrow Hindi. For example, the terminology used to describe parts of speech (POS) in Hindi Wordnet is completely foreign to Urdu speakers. The POS names are Sanskrit-based whereas Urdu POS words are Persian-based and Arabic-based. In Hindi the word for noun is sanya and in Urdu it is ism. The proper noun in Hindi is called Vyakti vachak sang. No Urdu speaker would know this unless they have studied Hindi grammar. In order to work through these differences, one has to be familiar with both languages at almost expert levels.

5. Conclusion and Remarks

In this position paper I contend that Hindi and Urdu are two different languages, at least for computational processes. The reasoning in this paper does not suggest that Urdu and Hindi are two different languages in all aspects, but rather asserts that the
current tools available for Hindi language cannot be used for Urdu language processing. In other words, in order to use Hindi resources to do Urdu computational processing, one has to know Hindi at a detailed linguistic level. The examples of the quantitative analysis confirm the emerging research from the Urdu and Hindi language researchers that there is trend of divergence between the two languages. There is a great opportunity to use casual middlebrow to bridge these languages and develop tools for multilingual Information Access. In the meantime, new resources need to be created to do research in Urdu Information Access.

6. REFERENCES
A Patient Support System based on Crosslingual IR and Semi-supervised Learning

Hideki Isozaki, Tsutomu Hirao, Katsuhito Sudoh, Jun Suzuki, Akinori Fujino, Hajime Tsukada, Masaaki Nagata
NTT Communication Science Laboratories
NTT Corporation
2-4 Seikacho, Sorakugun, Kyoto, Japan, 619-0237
{isozaki,hirao,sudoh,jun,a.fujino,tsukada,nagata}@cslab.kecl.ntt.co.jp

ABSTRACT
Even though patients are now using the Web to get useful information, the latest medical information is not available in most languages, except English. Even if patients want to learn about current treatments, they do not want to read English documents filled with technical terms. To mitigate this situation, we are building a patient support system that combines crosslingual information retrieval, machine translation, and technical term extraction to provide up-to-date medical information.

Categories and Subject Descriptors
I.2.7 [ARTIFICIAL INTELLIGENCE]: Natural Language Processing—Machine Translation
J.3 [Computer Applications]: LIFE AND MEDICAL SCIENCES—Medical Information Systems

General Terms
CLIR

Keywords
medical information system, crosslingual information retrieval

1. INTRODUCTION
Patients are using the Web more and more to get useful information (www.nytimes.com/2008/09/30/health/30on-line.html). English-speaking patients can obtain useful medical documents from such different sources as:

- Patient community sites (e.g., PatientsLikeMe) and patient blogs.
- Reports from public medical organizations (e.g., WHO, NIH, NCI, and FDA).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR Workshop on Information Access in a Multilingual World July 23, 2009 Boston, MA USA

However, the latest medical information is generally not available in languages other than English. To support patients who want to learn about up-to-date treatments but who do not want to read English documents filled with technical terms, we are building a patient support system.

Similar patient support services are already available on the Web. Nikkei BP’s Cancer Navi (cancernavi.nikkeibp.co.jp/) provides cancer news in Japanese. Cancer Information Japan (cancerinfo.tri-kobe.org/) translates the National Cancer Institute (NCI)’s Physician Data Query (PDQ) into Japanese. This site provides two versions of a translation: one for experts and the other for patients. The Japan Association of Medical Translation for Cancer (ww.cancerit.jp/xoops/) translates PubMed abstracts.

However, these services are labor intensive and expensive. Our motivation is to automate them as much as possible.

Data mining and natural language processing (NLP) of biomedical literature are crucial research topics (compbio.uch.ac.edu/BioNLP2009/index.shtml), and useful BioNLP tools are available from the National Centre for Text Mining in the UK (NaCTeM, www.nactem.ac.uk/), the Tsujii Laboratory of The University of Tokyo (www-tsujii.is.s.u-tokyo.ac.jp/), and so on.

Our project’s goal is to analyze not only academic documents (PubMed abstracts and full papers) that are relatively uniform and written for experts but also patient sites and reports form public/governmental organizations.

Even though patient sites are not always completely reliable, they are much easier to understand than papers written by doctors and sometimes provide useful information.

In the future, our system will provide a stereoscopic view: the views of medical experts and patients. The system will also integrate worldwide information (e.g., a certain medicine is available in the USA but not in Japan) and local information (e.g., hospital H’s doctor D is good at a certain surgical operation).

Information credibility is essential in such a system. The above services explicitly disavow endorsement and liability, and we will also follow a similar policy. However, information credibility on the web is a critical research issue (www.d1.kuis.kyoto-u.ac.jp/vico/2/), and we will introduce some credibility handling functions to our system.

2. CURRENT STATUS
The current system consists of the following modules:

- Dictionary-based crosslingual information retrieval module
When a patient enters a Japanese term, it is translated into corresponding English terms. We use Life Science Dictionary (LSD) (lds.pharm.kyoto-u.ac.jp/en/) for Japanese-to-English term translation and Indri (www.lemurproject.org/indri/) for document retrieval.

- **Technical term recognizers**
  Patients do not recognize the substance names from LSD, and the system adds formal names, and instead most medical articles use its substance name, and "oseltamivir" [All Fields] OR "oseltamivir" [MeSH Terms] OR "oseltamivir" [Other] OR "oseltamivir" [Other] OR "oseltamivir" [Other], and not in the METHOD part that describes the details of experiments.

These tags are useful in document retrieval and crosslingual multidocument summarization. For instance, we can skip the translation of the METHOD part for experiments but are eager to find effective treatments.

To cover other abstracts without these rhetorical tags, we implemented a semi-supervised rhetorical parser.

- **Hierarchical phrase-based statistical machine translator**[6]
  This module, which translates English documents into Japanese, is trained with a 20,000-sentence medical bilingual corpus on the Web and 2,150,000 dictionary entries. We are intensively working on improving this module and will report it elsewhere.

### 2.1 Crosslingual IR

When a Japanese patient enters “Tamiflu” in Japanese, the dictionary gives the English expression “Tamiflu.” However, “Tamiflu” is rarely used in PubMed because it is a product name, and instead most medical articles use its substance name: “oseltamivir phosphate” or “oseltamivir.”


We extracted correspondence between the product and substance names from LSD, and the system adds formal names to a complex query: “oseltamivir” [MeSH Terms] OR “oseltamivir” [All Fields] OR “tamiflu” [All Fields], yielding 963 items.

We also believe that patients want to know the trends of treatments. Therefore, the CLIR module analyzes the search engine’s output. First, the top documents are counted for each year. The module also counts the technical terms in the titles of the retrieved documents and shows their frequency and counts by year.

In Figure 1, the system found that gemcitabine appears 262 times in the titles of the top 300 documents when a patient entered “suigan” (pancreatic cancer) to the system. That is, almost all papers on pancreatic cancer have this medicine in their titles. When a patient enters “shinkei koushu” (glioma), the system shows that temozolomide appeared 92 times in the titles of the top 300 documents, and 48 times appeared from 2002 or later. That is, the popularity of treatment with this medicine is increasing.

In this way, document retrieval by disease name and the extraction of medicine names from titles reveals the trends of treatments. Frequent technical terms are shown with Japanese translations given by LSD, and they are linked to “Goo,” a Web search engine. Since retrieved documents are written in Japanese, patients can easily learn what gemcitabine is. In case of gemcitabine, Goo returns an entry of Japanese Wikipedia as the best document.

Patients can obtain useful information without reading poorly translated articles. This method can also be regarded as a redundancy-based factoid question answering system [2, 4, 1] that answers, “which medicine is used for this disease?” Since we can prepare a list of disease names, we can analyze the trend of treatments beforehand for each disease.

By clicking on one of the retrieved titles, we can see the entire text of the abstract. Fig. 2 shows the structure of the page. The left part of the page shows the original English abstract separated by a rhetorical parser described below. The right part shows the same abstract in Japanese. Medical terms detected by the technical term recognizers in the left part are colored. When the cursor is moved over them, their descriptions from different information sources pop up.

We need the English abstract because the Japanese translation is not very reliable yet. Japanese is completely different from English and even the state-of-the-art machine translators cannot generate satisfactory translation.

---

**Figure 1: Retrieval result page (italic: Japanese)**
Our goal is not word-to-word translation but comprehensible survey in Japanese. Our system will analyze English documents with information extraction/data mining technologies. Then, the system will generate a Japanese survey based on the analysis.

2.2 Semi-supervised learning

We applied semi-supervised learning to rhetorical parsing and technical term extraction.

2.2.1 Rhetorical parsing

Hirohata et al.’s rhetorical parsing [3] classifies each sentence in PubMed abstracts into four sections: OBJECTIVE, METHOD, RESULTS, and CONCLUSION. In addition, each section is further divided into two classes: beginning of the section (e.g., B-METHOD) and inside of the section (I-METHOD).

As features of each sentence, they used its relative location in the abstract and word unigrams and word bigrams selected with χ² criteria, etc. According to their experiments Conditional Random Fields (CRFs) outperformed Support Vector Machines (SVMs).

Following their experimental setting, we used 10,000 abstracts of their ‘pure’ corpus data for training. Table 1 shows that our semi-supervised learning method outperformed Hirohata et al.’s supervised CRF method, even though our implementation of the supervised CRF system has not achieved their performance level yet. Here, we used 100,000 abstracts as unlabeled data for semi-supervised learning.

This rhetorical structure is now used for selecting fields of document retrieval. For instance, such a complex query can be composed so that OBJECTIVE field has “pancreatic cancer” and the RESULTS field is “statistically significant” and so on. We are also planning to use such classification results for multidocument summarization.

2.2.2 Medical Term Recognition

Medical term recognition can be regarded as an extension of named entity recognition, which is popular in the Natural Language Processing community. Conventional named entity recognizers also use supervised learning approaches such as CRFs and SVMs. In such approaches, we do not have to write down complicated rules to detect named entities. But we have to prepare a large amount of training data. Since their performance is almost saturated, we employed a semi-supervised approach [5] that is robust against new terms.

We trained a semi-supervised CRF using the Penn BioIE Corpus (CYP450) bioie.ldc.upenn.edu/publications/latest_release/. We have just started a preliminary experiment with this corpus. When we used 73,108 sentences for training and 8,137 sentences for the tests, supervised CRFs gave F-measure = 0.897, which the semi-supervised CRFs slightly improved to 0.905 when we used the above 100,000 PubMed abstracts as unlabeled data. We do not know any experimental reports on this dataset from other researchers.

Since CYP450 covers only articles on cytochrome P450 enzymes, this tagger is not general enough. Therefore, we also built a simple left-to-right longest match tagger based on LSD.

3. CONCLUSION AND FUTURE WORK

We are building a CLIR system for patients who want to learn up-to-date treatments but who do not want to read English documents filled with technical terms. We showed that a simple combination of CLIR and medical term recognizers provides the trends of treatments. The system currently only covers PubMed entries, which are uniform, reliable, and relatively easy to analyze. Future work includes the analysis of patient and public organization sites. The introduction of patient sites will require judgments of credibility.

4. REFERENCES