Multilinguales Web Retrieval im Rahmen von WebCLEF 2006

Ben Heuwing, Robert Strötgen
Information Science, University of Hildesheim,
Marienburger Platz 22
D-31141 Hildesheim, Germany
stroetgen@uni-hildesheim.de

Abstract

Dieser Beitrag beschreibt Retrievalexperimente mit einem umfangreichen multilingualen Korpus im Rahmen von WebCLEF 2006 an der Universität Hildesheim. Im Vordergrund stand die Nutzung von HTML Strukturelementen, der Einsatz von Blind Relevance Feedback und die Evaluierung des sprachunabhängigen Indexierungsansatzes.

1 Einleitung

Einen umfangreichen, heterogenen und multilingualen Korpus effizient in Hinblick auf Rechenleistung zu indexieren und eine hohe Retrievalqualität zu erreichen, war Ziel der diesjährigen Experimente mit dem EuroGOV-Korpus im Rahmen des Web Tracks des Cross Language Evaluation Forum (CLEF).

Für WebCLEF 2005 konnte die Universität Hildesheim mit einem sprachunabhängigen Indexierungsansatz das beste System für mehrsprachiges Retrieval entwickeln [Jensen *et al.*, 2006]. In diesem Jahr sollte versucht werden, das bei der ersten Teilnahme im letzten Jahr verwendete System in Bezug auf die Vorverarbeitung und Säuberung der Korpusdaten zu verbessern. Ein weiteres Ziel war die Implementierung einer Blind Relevance Feedback Option und deren Evaluierung. Die direkte Herangehensweise mit einem multilingualen Index und ohne Übersetzung der Anfragen, die im letzten Jahr zum Erfolg vor allem beim multilingualen Task (Anfragen und Ergebnisse in verschiedenen Sprachen) geführt hatte, wurde erweitert, wobei die jeweils besten Ansätze (vor allem die Indexierung ganzer Wörter statt eines N-Gram Ansatzes) weiter verfolgt wurden. Da sich der multilinguale Track jedoch für viele Teilnehmer als wenig erfolgversprechend erwiesen hatte, wurde in diesem Jahr nur der Mixed-Monolingual Task angeboten, bei dem Ergebnisse nur in der Sprache der Anfrage gefordert sind. Zu Testzwecken wurde trotzdem mit dem neuen System ein multilingualer Run erstellt und eingereicht.

2 Das System

Das System bringt zunächst den Korpus in ein gut weiterzuverarbeitendes Format. Die Indexierung und die Suche wurde in Java und auf der Basis der Suchmaschinen-API *Apache Lucene* implementiert.

2.1 Vorverarbeitung des EuroGOV2-Korpus

Der 80GB große EuroGOV2-Korpus umfasst ca. 3.6 Mio. Internetseiten in verschiedenen Formaten in 20 verschiedenen Sprachen von den Seiten von Regierungen europäischer Länder und der EU. Der Korpus besteht aus 157 Dateien in einem XML-ähnlichen Format mit jeweils maximal 25.000 Dokumenten. Zu jedem Dokument gibt es Metadaten wie die Ursprungs-URL und Angaben aus dem HTTP-Header.

Die eigentlichen Inhalte der Dokumente befinden sich in einem CDATA-Bereich, innerhalb dieser Elemente werden XML-ähnliche Konstruktionen von einem XML-Parser nicht als solche aufgefasst. Allerdings konnten so verschachtelte CDATA-Elemente entstehen, die in XML nicht zulässig sind. Dies ist einer der Gründe, warum das Format des Korpus kein wohlgeformtes XML ist.

Bei der Vorverarbeitung werden daher nun zunächst alle nicht XML-konformen Zeichen [XML W3C Recommendation] heraus gefiltert. Realisiert wurde dies durch einen effizient auf Ebene des Zeichenstroms arbeitenden Filter. Fehlerquelle sind vermutlich die unterschiedlichen Zeichenkodierungen der Dokumente. Dies dürfte Auswirkungen auf die Retrievalqualität vor allem bei problematischen Kodierungen haben.

Einige nicht maskierte Sonderzeichen der in den Metadaten angegebenen URLs müssen im nächsten Schritt unter der Verwendung von regulären Ausdrücken behandelt werden. Auf diese Weise werden auch verschachtelte CDATA- Elemente entfernt, die in wohlgeformten XML-Dokumenten nicht zugelassen sind.

Um mittels eines XML-Parsers gezielt auf Inhalte von einzelnen in den Dokumenten enthaltenen HTML-Elementen zuzugreifen, können diese aus dem CDATA-Bereich extrahiert und als neue Elemente in die XML-Struktur eingefügt werden. Aufgrund dieser Maßnahmen konnten alle Text- bzw. HTML-Dokumente indexiert werden.

2.2 Indexierung

Das entwickelte System setzt als Basis Suchmaschine *Lucene* [Lucene Projekt Homepage] ein. *Lucene* erlaubt die Erstellung von Indizes mit mehreren Feldern. Aufgrund einer vorherigen Analyse der Häufigkeit des Auftretens der verschiedenen HTML-Elemente wurde entschieden, die Inhalte von `<title>` und `<h1>`-Elementen zu einem Indexfeld *title*, die Inhalte der anderen extrahierten Elemente in einem Feld *emphasised* zusammenzufassen, und so beim Retrieval mit unterschiedlichen Gewichtungen experimentieren zu können. Das Dokument wird einmal komplett (*content*) und einmal beschnitten auf 50 Wörter aus der Mitte des Dokuments (*content_cutoff*) indiziert. Termvektoren für das Blind Relevance Feedback (BRF) werden nur für die Felder *title*, *emphasised* und *content_cutoff* berechnet. Auf die Nutzung der gesamten Volltexte für das BRF wurde mit Rücksicht auf die be-

schränkten Ressourcen vor allem bezüglich der Indexerstellung verzichtet. Der Index mit Termvektoren hatte eine Größe von ungefähr 6GB.
Die bereits vorhandene multilinguale Stopwortliste, die 13 Sprachen umfasst, wurde um die im Korpus am häufigsten auftretenden Wörter erweitert. Diese erweiterte Liste beinhaltete 4722 Wörter. Die Idee, eine weitere, titelspezifische Liste einzusetzen, die z.B. auch automatisch erstellte und daher nicht bedeutungstragende Konstruktionen wie ‚no title' berücksichtigt, wurde fallen gelassen, da diese nicht in dem hohen Maße wie vermutet auftraten.

Die Einteilung in einzelne Tokens wurde dem in Lucene vorhandenen *StandardAnalyzer* überlassen, dabei werden bei den Wörtern -s Endungen entfernt und Interpunktion behandelt. Im letzten Jahr hatte sich gezeigt, dass diese Methode auch sprachunabhängig gute Ergebnisse liefern kann.

2.3 Retrieval

Um die Anfragen zu erstellen, wird der *Lucene QueryParser* eingesetzt. Die daraus entstandene Anfrage kann dann durch Gewichtung und Blind Relevance Feedback modifiziert werden. Das Ranking der Ergebnisse basiert auf einer längennormalisierten tf-idf-Formel.

Vor dem Hintergrund des Mixed-Monolingual Tasks nutzen wir zusätzlich die in den Metadaten der Topics angegebenen Zieldomains, um die Ergebnisse auf Dokumente in dieser Domain einzuschränken. Dies ist ohne großen Rechenaufwand möglich durch die *QueryFilter*-Klasse, die den zusätzlichen Vorteil hat, dass sie die Rankingergebnisse nicht weiter beeinflusst. Die Domains sind Teil der Dokument-IDs und werden während des Indexierens extrahiert und in ein eigenes Feld geschrieben. Ohne diesen Schritt musste eine Suche mit Platzhaltern auf den IDs ausgeführt werden, was zu Performance-Problemen führte.

Es wurde darauf geachtet, dass die einzelnen Felder beliebig gewichtet werden konnten, diese Option ist zu Evaluierungszwecken auch über die Kommandozeile verfügbar. Eine hohe Gewichtung des *title*-Feldes (20:1:1 im Verhältnis zu den beiden anderen Inhaltsfeldern) brachte hierbei die größten Vorteile.

Ebenfalls zu Evaluierungszwecken kann die Anzahl der für das Blind Relevance Feedback eingesetzten Dokumente und der für die Anfrage verwendeten Terme sowie die Gewichtung zur ursprünglichen Anfrage von der Kommandozeile aus verändert werden. Es wird eine an der Universität Hildesheim für den CLEF Ad-Hoc Track erstellte Implementierung eingesetzt, die auf den von Lucene bereitgestellten Termvektoren arbeitet und die Termgewichte über einen Robertson Selection Value berechnet [Hackl *et al.,* 2005].

3 Ergebnisse

3.1 Topics WebCLEF 2006

Wie schon erwähnt stand in diesem Jahr der Mixed-Monolingual Task im Vordergrund [WebCLEF Homepage]. Die Topics für dieses Jahr bestanden aus 319 manuell erstellten Topics (neu erstellte und ein Teil der Topics von WebCLEF 2005) und 1620 automatisch erstellten, das hierzu verwendete Verfahren wurde nicht bekannt gegeben.

3.2 Runs

Für die Teilnahme wurden nach Experimenten mit den Topics von WebCLEF 2005 ein Run mit starker Gewichtung des Titels und zwei Runs mit unterschiedlich gewichtetem Blind Relevance Feedback zur Anfrageerweiterung sowie ein Run für den multilingualen Task erstellt. Für alle Runs wurde zum Vergleich der Mean Reciprocal Rank[1] (MRR) mit den Topics von WebCLEF 2005 berechnet und die durchschnittliche Häufigkeit eines Treffers unter den ersten 5, 10, 20, 50 Ergebnissen (*average success*) ermittelt.

Die Runs für den Mixed-Monolingual Task zeigen in den Experimenten (Tabelle 1) insgesamt verbesserte Ergebnisse im Vergleich zu denen von WebCLEF 2005 und den danach durchgeführten Postexperimenten (Tabelle 2). Der beste eingereichte Run für den Mixed-Monolingual Task von WebCLEF 2005 erreichte einen MRR von 0,1603, während in den Postexperimenten ein MRR von 0,2377 erreicht wurde [Jensen *et al.*, 2006]. Der beste Run (*UHiTitle*) in diesem Jahr zeigte mit einem MRR von 0,2807 also eine Verbesserung von 0,043 Punkten. Der *average success at 50* verbesserte sich von 0,253 auf 0,5192.

Die Verbesserungen können auf die umfassendere Indexierung und die erweiterte Stopwortliste, vor allem aber auf die Verwendung zusätzlicher Metadaten (Zieldomain) zurückgeführt werden. Vermutlich führte die umfassendere Indexierung zu einem höheren Recall (Auswirkungen auf *average success*) und der Domainfilter vor allem zu höherer Precision (Auswirkungen auf MRR). Ein ansonsten dem diesjährigen Run *UHiTitle* entsprechender Durchlauf ohne Domainfilter zeigte weniger deutliche Verbesserungen beim MRR (0,0175 Punkte), aber trotzdem einen deutlich verbesserten *average success at 50* von 0,4570.

Der multilinguale Run hatte im Vergleich zum besten des letzten Jahres einen nur leicht verbesserten MRR von 0.2134, der *average success at 50* verbesserte sich dagegen deutlich auf 0.4004. Hier wurde kein Domainfilter verwendet, die umfassendere Indexierung hatte jedoch wahrscheinlich positive Auswirkungen auf den Recall.

Das Blind Relevance Feedback in der durchgeführten Form scheint nicht zu Verbesserungen geführt zu haben. Allerdings kann der Einsatz eventuell durch weitere Experimente optimiert werden.

Name	Bemerkungen	MRR	Average Success at 50
UHiTitle	Titel^20	0.2807	0.5192
UHiBRF1	BRF (Gewichtung 1.0)	0.2731	0.4973
UHiBRF2	BRF (Gewichtung 0.5)	0.2771	0.5082
UHiMu	multilingual	0.2134	0.4004

Tabelle 1: Ergebnisse der Runs von 2006 mit Topics von 2005

Name	Bemerkungen	MRR	Average Success at 50
UHiSMo	offizieller Mixed-Monolingual Run	0.1603	0.2870
UHiSMu	offizieller multilingual Run	0.1370	0.2587
UHiSTiMo	bester Mixed-Monolingual der Postexperimente	0.2377	0.2530
UHiSTi101	multilingual Run aus Postexperimenten	0.2117	0.2117

Tabelle 2: Ergebnisse WebCLEF 2005 und Postexperimente [Jensen *et al.*, 2006]

[1]MRR = 1 / Rang des ersten relevanten Dokuments in der Ergebnisliste [Jensen *et al.*, 2006]

Literatur

[Hackl *et al.,* 2005] Hackl, René; Mandl, Thomas; Womser-Hacker, Christa (2005): Ad-hoc Mono- and Multilingual Retrieval Experiments at the University of Hildesheim. In: Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Sept. 2005, Wien.

[Jensen *et al.*, 2006] Niels Jensen, René Hackl, Thomas Mandl, Robert Strötgen. Web Retrieval Experiments with the Euro-GOV Corpus at the University of Hildesheim. In: Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers. Springer [LNCS 4022]

[Lucene Project Homepage] http://lucene.apache.org

[WebCLEF Homepage] http://ilps.science.uva.nl/WebCLEF/

[XML W3C Recommendation] http://www.w3.org/TR/2004/REC-xml11-20040204/#charsets