Thomas Mandl, Christa Womser-Hacker

Eigenschaften von natürlichsprachlichen Topics in Information Retrieval Experimenten

Zusammenfassung: In empirischen Evaluierungsstudien im Information Retrieval stellt sich immer wieder Frage nach der Validität der Ergebnisse. Eine Analyse der Topics des Cross Language Evaluation Forum (CLEF) sollte zeigen, ob linguistische Eigenschaften der Formulierungen der Topics Rückschlüsse auf die Qualität von Retrievalergebnisse zulassen. Für die englischen Topics konnten nur schwache Korrelationen zwischen durchschnittlicher Retrievalqualität und der Gesamtanzahl linguistischer Phänomene nachgewiesen werden. Für die Kombination von Retrievalsystemen, ihren Eigenschaften und den Eigenschaften den CLEF-Topics konnten keine Beziehungen gefunden werden.

1 Information Retrieval in mehrsprachigen Kontexten

Information Retrieval (IR) beschäftigt sich mit der Suche nach Information und mit der Repräsentation, Speicherung und Organisation von Wissen. Information Retrieval modelliert Informationsprozesse, in denen Benutzer aus einer großen Menge von Wissen die für ihre Problemstellung relevante Teilmenge herauslösen. Dabei entsteht Information, die im Gegensatz zum gespeicherten Wissen problembezogen und an den Kontext angepasst ist.

IR gewinnt im Zeitalter des Internet neue Bedeutung (cf. Baeza-Yates & Ribeiro-Neto 1999). Der großen Menge gespeicherten und online zugänglichen Wissens stehen zahlreiche frei nutzbare Internet-Suchmaschinen gegenüber. Damit steigt auch der Bedarf für die Evaluierung von IR-Systemen. Die Evaluierung, die seit den 60er Jahren ihr Methodeninventar sukzessive verfeinert hat, steht angesichts der neuen Möglichkeiten vor neuen Herausforderungen.

1.1 Verfahren für mehrsprachiges Retrieval

Mehrsprachiges Information Retrieval oder Cross Language Information Retrieval (CLIR) geht von der Annahme aus, dass Benutzer eine Fremdsprache zwar häufig passiv beherrschen und die Relevanz von Dokumenten in dieser Sprache zumindest abschätzen können, dass sich aber Probleme bei der Erstellung von Anfragen ergeben können. In diesem Benutzungskontext entsteht ein Mehrwert, wenn eine einsprachige Anfrage zu Dokumenten in mehreren Sprachen führt und diese ausschließlich nach

Relevanz anordnet. Wie bei anderen Formen der semantischen Heterogenität (cf. Mandl 2001) erfordern mehrsprachige Information Retrieval Systeme einen Transformationsprozess. Einen Überblick über die Funktionalität eines CLIR-Systems zeigt Abbildung 1.

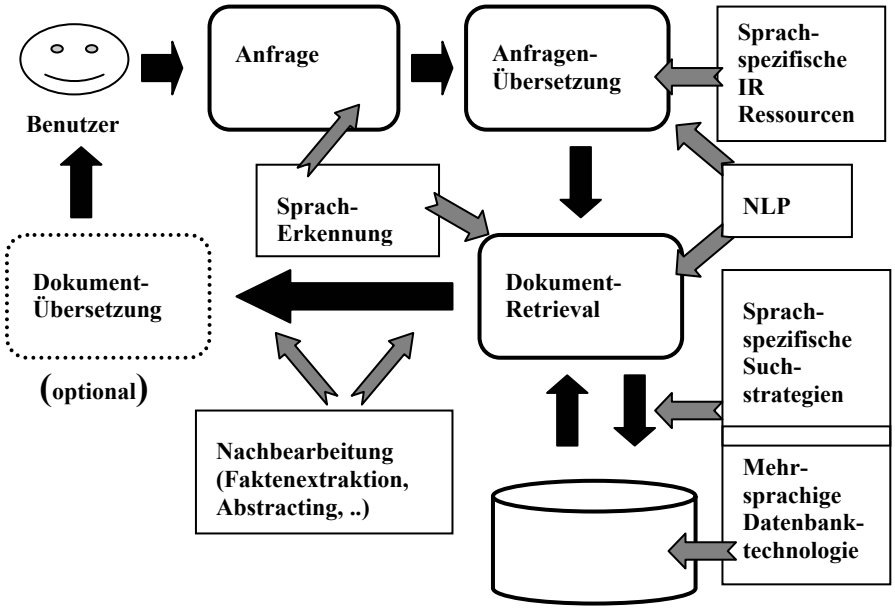

Abb. 1: CLIR im Überblick (nach Evans et al. 2002).

Beim mehrsprachigen IR treten neben allgemeine IR-Probleme die vielfältigen Aspekte der Übersetzung (einschließlich der jeweils vorhandenen linguistischen und lexikalischen Hilfsmittel für die verschiedenen Sprachen) und der integrierten Ausgabe der Ergebnisse aus mehreren Dokumentmengen hinzu.

Crosslinguales IR (CLIR) versucht, auf eine Anfrage in einer Sprache Dokumente in anderen Sprachen zu liefern und sucht in einem multilingualen Korpus nach relevanten Dokumenten. Für diese komplexe Aufgabe haben sich Fusionsverfahren etabliert, die auf mehreren Ebenen Evidenzen aus mehreren Quellen kombinieren (Savoy 2002).

1.2 Internationale Evaluierungsinitiativen

Seit Beginn der Evaluierung im IR setzten Forscher für ihre Experimente unterschiedliche Testkollektionen ein. Daher waren ihre Ergebnisse kaum vergleichbar. Verschiedene Initiativen stellen in den letzten Jahren standardisierte Kollektionen zur Verfügung und haben so die Vergleichbarkeit zwischen den Systemen verbessert. Seit drei Jahren entwickelt das Cross-Language Evaluation Forum (CLEF[1], cf. Kluck et al. 2002) Methoden und eine Infrastruktur für die Bewertung sprachübergreifender Suchverfahren. CLEF basiert auf Erfahrungen aus der amerikanischen TREC-Initiative[2] (Text Retrieval Conference, cf. Voorhees & Harman 2001). Parallel entstand in Japan das NTCIR[3] Projekt für multilinguales Retrieval in asiatischen Sprachen (cf. Kando et al. 2001).

In den USA begann 1989 das National Institute of Standards and Technology (NIST) in Gaithersburg (Maryland) mit einem Projekt zur Bewertung von IR-Systemen. Unter der Leitung von Donna Harman stellt das NIST umfangreiche Textkollektionen, Benutzerbedürfnisse in Form von sog. Topics und die Infrastruktur für die Evaluierung zur Verfügung. Auf der jährlichen Text Retrieval Conference (TREC) stellen die Teilnehmer ihre Systeme und Ergebnisse vor. Die Initiative findet großen Anklang: so beteiligten sich an TREC 2001 bereits 86 Forschungsgruppen aus Industrie und Wissenschaft mit ihren Systemen. Die Ergebnisse und die Artikel der Teilnehmer stehen online zu Verfügung.

CLEF führt den von TREC eingeführten Cross-Language Track für europäische Sprachen fort und orientiert sich dabei weitgehend am Ablauf des Ad-hoc-Retrieval Track, während sich TREC-CLIR auf Sprachen des arabischen Sprachraums konzentriert. Im CLEF-Projekt arbeiten Gruppen aus verschiedenen europäischen Ländern (und damit auch Sprachräumen) mit dem NIST zusammen.[4]

Die CLEF-Organisatoren erstellen Topics für Testfragestellungen in drei Detaillierungsebenen. Neben einer aus wenigen Worten bestehenden Überschrift (Title) und einer Kurzbeschreibung (Description) des Themas in einem Satz gibt es eine ausführliche Beschreibung (Narrative). Die Teilnehmer entscheiden sich für eine der Fassungen oder eine Kombination (z.B. Überschrift und Kurzbeschreibung oder alle drei Elemente) und arbeiten damit. Die besondere Problematik der asiatischen Sprachen, die bereits auf der Ebene der Zeichen ganz andere Anforderungen stellen als europäische Sprachen, führte schon früh zu speziellen IR-Verfahren. Seit 1997 widmet sich dieser Thematik mit NTCIR auch ein eigenes Evaluierungsprojekt, das ähnlich wie

---

1 http://www.clef-campaign.org
2 http://trec.nist.gov
3 http://research.nii.ac.jp/ntcir/
4 IEI-CNR (Pisa, Italien) als Koordinator, Eurospider (Zürich, Schweiz), ELRA (Paris, Frankreich), IZ (Bonn, Deutschland), UNED (Madrid, Spanien), NIST (Gaithersburg, USA).

TREC und CLEF Korpora und Aufgabenstellungen entwickelt und die Bewertung übernimmt. Der zweite NTCIR-Workshop fand 2001 statt und konzentrierte sich auf Chinesisch und Japanisch.

2 Validität von Retrieval-Experimenten

Empirische Untersuchungen im Information Retrieval erfordern seit jeher einen erheblichen Aufwand. Um diesen zu rechtfertigen, müssen die Ergebnisse aussagekräftig, möglichst signifikant und valide sein. Die Meta-Ebene von Information Retrieval Experimenten war daher seit jeher ein wichtiger Forschungsgegenstand. Untersucht wurden etwa die statistische Signifikanz der Ergebnisse (Tague-Sutcliffe & Blustein 1993), die Konsistenz der Relevanz-Bewertungen durch Juroren und nicht zuletzt die Qualität der Aufgabenstellungen.

2.1 Untersuchung der Relevanz

Die Subjektivität der Relevanz-Bewertungen von Juroren hat bereits mehrfach zu Zweifeln an der Zuverlässigkeit von Experimenten geführt. Binäre Relevanz-Urteile sind offensichtlich subjektiv geprägt und trotz aller Richtlinien lässt sich keine Vereinheitlichung der Maßstäbe herbeiführen. Da diese Urteile die Basis der Ergebnisse liefern, könnte die Subjektivität die Ergebnisse verfälschen. Eine aktuelle Studie bestätigt zwar die Subjektivität der Urteile, zeigt aber, dass die Folgerung nicht zutrifft. Bei der Untersuchung wurden für mehrere Topics der TREC-Initiative zusätzliche Relevanz-Urteile von unterschiedlichen Juroren erhoben. Es zeigte sich, dass diese tatsächlich unterschiedlicher Meinung über die Relevanz waren. Allerdings wirkte sich dies nicht auf die Reihenfolge der Systeme aus. Zwar war die absolute Qualität der Systeme abhängig vom Juror unterschiedlich, allerdings zielt TREC auf ein Ranking der Systeme ab, um vergleichende Aussagen treffen zu können. Die Reihenfolge blieb weitgehend unverändert (Voorhees 1998). Solange also eine Person konsequent ihren Standpunkt auf die Ergebnis-Dokumente anwendet, ergeben sich keine Verfälschungen im Endergebnis. In CLEF mag sich dies anders darstellen, da hier Muttersprachler die Ergebnisse bewerten und somit bei Listen mit mehreren Sprachen unterschiedliche individuelle Standpunkte eingehen. Eine entsprechende Untersuchung der Auswirkungen steht für CLEF noch aus.

2.2 Aufgaben für die Information Retrieval Evaluierung

Die Anfragen für die Evaluierung von Information Retrieval Systemen drücken ein Informationsbedürfnis aus und sind in natürlicher Sprache formuliert. Besonders Verfälschungen durch die Formulierung der Anfragen sollten vermieden werden. In TREC und CLEF wird versucht, die Topics möglichst auf natürliche Informationsbedürfnisse zu gründen. Das folgende Beispiel soll dies illustrieren:

<num> C007

<S-title>

Consumo de drogas y el fútbol

<S-desc>

Encontrar documentos sobre el consumo de drogas en el fútbol.

<S-narr>

Los documentos relevantes informan sobre casos de jugadores de fútbol condenados por el consumo de drogas. Las discusiones generales sobre temas relacionados con la droga en el mundo del fútbol también son relevantes.

</top>

Im Information Retrieval wurde immer wieder betont, dass die Qualität der Aufgabenstellungen entscheidend für die Testergebnisse ist („the quality of requests (and hence queries) appears very important“, Sparck Jones 1995).

Auch bei den Aufgabenstellungen kann durch die Übersetzung eine Modifikation des Inhalts eintreten, so dass ein ausgewogener Prozess entwickelt wurde, in dem die Topics in einem mehrstufigen Verfahren übersetzt und überprüft werden (Kluck & Womser-Hacker 2002). Besonders kulturelle Eigenheiten sollten so weit wie möglich erkannt werden (Womser-Hacker 2002). Hier kann es übrigens bereits innerhalb einer Sprache zu lexikalischen Problemen kommen, wenn unterschiedliche Kulturräume beteiligt sind. Die CLEF Dokumente umfassen sowohl Texte aus Deutschland als auch der Schweiz, so dass Varianten wie *Abschiebung* und *Ausschaffung* berücksichtigt werden müssen, um eine gleiche Ausgangsbasis der Retrieval-Systeme zu gewährleisten.

Dieser Beitrag untersucht weitergehende sprachliche Eigenschaften der Topic-Formulierungen und analysiert, inwieweit sich diese auf die Qualität der System-Antworten auswirken und ob eventuell einzelne Systeme mit bestimmten linguistischen Phänomenen besser zurechtkommen und demnach qualitativ höherwertigere Resultate liefern. Dieser Beitrag zeigt somit beispielhaft, wie sich formale Untersuchungen sprachlicher Eigenschaften im Information Retrieval einsetzen lassen.

Damit wird auch in eine zweite Forschungsrichtung verwiesen. Während die genannten Abhängigkeiten für eine Evaluierungsstudie als Nachteil gelten, bilden Eigenschaften der Anfragen seit längerem einen Ansatzpunkt für die Verbesserung von Retrieval-Systemen. Je nach sprachlichen Phänomenen in der Anfrage können unterschiedliche System-Parameter gewählt werden (Mandl & Womser-Hacker 2000). Beispiele bieten Kwok & Chan 1998 und Wilkinson et al. 1995, die ihr Retrieval-System abhängig von der Länge der Anfrage anders parametrisieren.

3 Topics von CLEF 2001

Wie in den vorhergehenden Jahren, wurden die Topics der CLEF 2001 Kampagne nicht konstruiert (etwa auf der Basis von Dokumenten), sondern drücken in möglichst natürlicher Weise ein potentielles Informationsbedürfnis aus. Unsere Untersuchung fokussiert auf die folgenden Fragen:

- Was sind die wichtigsten sprachlichen Eigenschaften der CLEF Topics?
- Haben diese Eigenschaften Einfluss auf die Qualität des Retrievals?
- Können derartige Erkenntnisse eventuell für die Verbesserung von Systemen ausgenutzt werden?
- Sollten die CLEF Topics verändert werden, um Einflüsse durch bestimmte sprachliche Eigenschaften zu vermeiden?

Eine detaillierte Betrachtung der CLEF-Ergebnisse führt zwangsläufig zu diesen Fragestellungen. Die Ergebnisse der einzelnen Systeme weisen auffällige Streuungen für die unterschiedlichen Topics auf. Dies soll zunächst an einem Beispiel illustriert werden, worauf eine intellektuelle Analyse folgt. Den Kern der Untersuchung bildet aber die automatische Abbildung von linguistischen Eigenschaften der Topics auf die Qualität der Ergebnisse.

Für einzelne Topics ergeben sich Auffälligkeiten bei den Retrievalergebnissen. Zum Beispiel weist das System EIT01M3N in CLEF 2001 eine relativ gute durchschnittliche Precision von 0,341 auf. Das bedeutet, das System hat über alle 50 Topics durchschnittlich diese Qualität erreicht. Für Topic 44 allerdings führt das System nur zu einer Precision von 0,07, obwohl dieses Topic lediglich durchschnittliche Schwierigkeit aufweist und der Durchschnitt aller Systeme bzw. Runs bei 0,27 liegt. Eine intellektuelle Analyse der Topics zeigt, dass zwei der schwierigsten Topics keine Eigennamen enthalten und von Sport handeln (Topic 51 und 54).

	Durch-schnitt	Std. - Abweichung	Maximum	Minimum
Alle Runs	0,273	0,111	0,450	0,013
Topics über alle Sprachen	0,273	0,144	0,576	0,018
Englische Runs	0,263	0,074	0,373	0,104
Englische Topics	0,263	0,142	0,544	0,018
Deutsche Runs	0,263	0,092	0,390	0,095
Deutsche Topics	0,263	0,142	0,612	0,005

Tabelle 1: Überblick über die Ergebnisse.

Im Detail interessieren linguistische Charakteristika der CLEF Topics, welche Hinweise auf die Performanz der Systeme geben. Information Retrieval Systeme beinhalten linguistische Komponenten für Mehrwortanalysen oder Grundformreduktion sowie Heuristiken wie Stoppwortlisten und Eigennamenerkennung. Sprachliche Phänomene stellen demnach Herausforderungen für Systeme dar. Einzelne Systeme könnten nun besser auf bestimmte Herausforderungen in Anfragen abgestimmt sein, während andere damit Probleme haben. Solche Untersuchungen sind auch deshalb sinnvoll, weil die Abweichung der Qualität zwischen den Topics meist höher ist als zwischen den Systemen (cf. Womser-Hacker 1997).

Tabelle 1 zeigt den Durchschnitt der Qualität aller Systeme. Dieser unterscheidet sich überraschenderweise nicht für Deutsch und Englisch als Topic-Sprache. Für die weitere Analyse betrachten wir sowohl die Performanz für alle Topics (alle Systeme für ein Topic) als auch die Performanz für alle Systeme (ein System für alle Topics). Hier liegt sowohl für Deutsch als auch für Englisch die Abweichung für die Topics höher als die für die Systeme (Runs). (>0,14 gegenüber <0,12).

Allerdings lassen sich keine allgemein „leichten“ Topics finden, die alle Systeme gut lösen, noch sind einzelne Systeme immer überlegen. Vielmehr liegt die Korrelation zwischen der durchschnittlichen Precision und der Abweichung für ein Topics über alle Systeme bei 0,83. Das bedeutet, je „leichter“ ein Topic ist, desto höher ist die Varianz der Runs für dieses Topic. Also erzielen nicht alle Systeme die gleiche Retrieval-Qualität für solche einfachen Topics.

Ebenso ist die Korrelation zwischen durchschnittlicher Precision und der Abweichung für einen Run über alle Topics eher hoch und beträgt 0,84. Daraus lässt sich schließen, dass kein System gut für alle Topics abschneidet. Im Gegenteil, je besser ein System, desto höher wird seine Abweichung von der durchschnittlichen Retrieval-Qualität bei den einzelnen Topics.

Daraus lässt sich die Hypothese ableiten, dass linguistische Auffälligkeiten wie morphologische Variationen, Abkürzungen oder Eigennamen Herausforderungen für

Retrieval-Systeme darstellen, welche diese unterschiedlich handhaben. Sehr gute Systeme könnten z.B. Probleme bei Anfragen mit Abkürzungen haben.

Ließen sich solche Abhängigkeiten entdecken, so könnten die Systeme automatisch darauf reagieren und die Topics jeweils an Systeme weiterleiten, welche die jeweiligen Phänomene gut behandeln. Bei einer intellektuellen Analyse wurden folgende Eigenschaften der Topics erfasst, wobei jeweils Types sowie Tokens aufgezeichnet wurden.

Deutsche Topics:	**Englische Topics:**
– Ursprüngliche Topic Sprache – Länge – Komposita – Abkürzungen – Nominalphrasen – Eigennamen – Negationen – Untergeordnete Sätze – Fremdwörter – Zahlen oder Daten	– Ursprüngliche Topic Sprache – Länge – Abkürzungen – Nominalphrasen – Eigennamen – Negationen – Untergeordnete Sätze – Fremdwörter – Klammerungen – Zahlen oder Daten

Tabelle 2: Topic Eigenschaften.

Zunächst wurden Korrelationen zwischen Eigenschaften der Topics und der durchschnittlichen Schwierigkeit betrachtet. Die Qualität eines Systems für ein Topic lässt sich mit folgenden Parametern messen:

- Durchschnittliche Precision für alle relevanten Dokumente
- Precision nach fünf Dokumenten
- Für einige Analysen wurden die Runs in fünf Qualitätsgruppen eingeordnet, so dass jeweils 20% aller Runs in einem Cluster lagen

Zwischen den einzelnen in Tabelle 2 aufgeführten Eigenschaften und der Qualität ergaben sich kaum nennenswerte Korrelationen, deren Betrag den Wert von 0,2 überstiegen hätte. Lediglich für Eigennamen wurde ein höherer Wert erreicht. Eigennamen scheinen ein Topic also „leichter“ zu machen.

	Eigennamen Types	Eigennamen Tokens	Summe aller linguistischen Phänomene
Englisch	0,446	0,473	0,293
Deutsch	0,440	0,464	0,286

Tabelle 3: Korrelation zwischen der Anzahl von Eigennamen und der Precision Überblick über die Ergebnisse.

Ebenso liegt für die Summe aller erfassten Phänomene eine Korrelation zur Qualität vor. Besonders wenn die Runs in Qualitäts-Gruppen eingeteilt werden, liegt die Korrelation bei 0,96 für die englischen Topics. Diese Beziehung wird in Abbildung 2 deutlich.

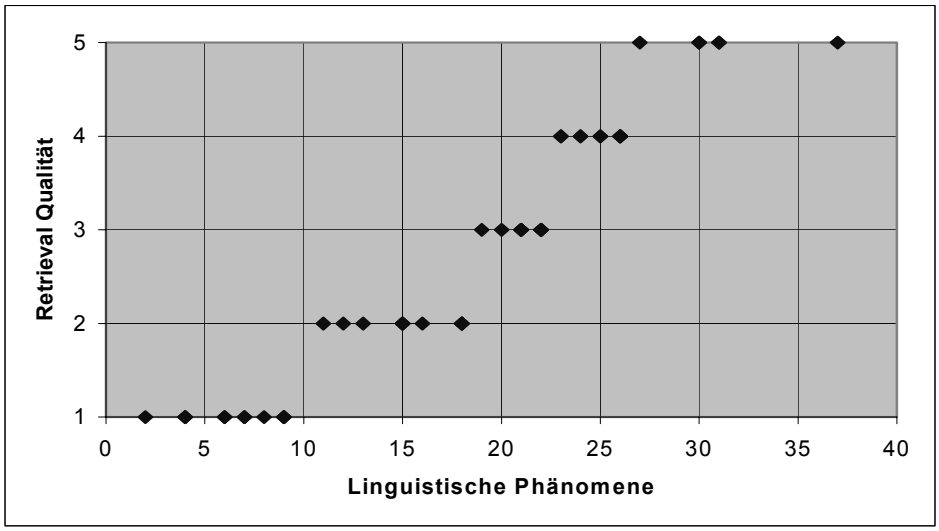

Abb. 2: Beziehung zwischen der Anzahl der linguistischen Phänomene und der Qualität eines Runs (1=schlecht, 5=sehr gut).

4 Topics und Runs in CLEF 2001

Die globale Analyse wurde durch eine Untersuchung der Systeme ergänzt, die den Zusammenhang zwischen einzelnem System und Topic Eigenschaften betrachtet. Dazu wurden folgende Eigenschaften der Runs aus dem CLEF-Material erhoben:

- Multi- oder bilingualer Run (zielt das Retrieval von einer Anfragesprache lediglich auf Dokumente in einer zweien Sprache oder auf Dokumente in mehreren Sprachen)
- Topic Sprache (Sprache, in der ein Topic zuerst formuliert wurde)
- Vom System benutzte Felder des Topics (title, description, narrative)
- Für den Auswertungspool benutzt oder nicht (nur Dokumente im Pool werden intellektuell auf Relevanz bewertet)

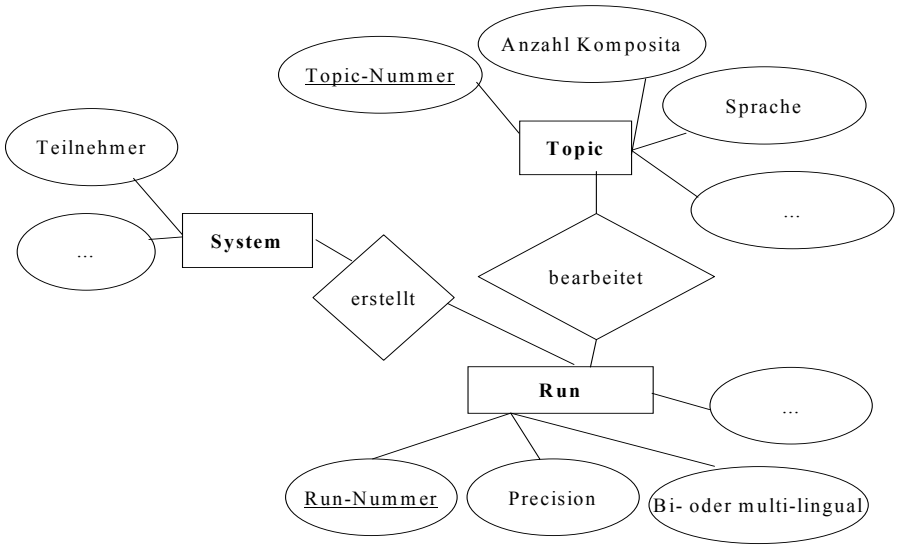

Abb. 3: Datenmodell für Topics, Systeme und Runs.

Das grundlegende Datenmodell für die beteiligten Objekte zeigt Abbildung 3. Eine statistische Analyse bot sich nicht an, da zu wenig Anhaltspunkte vorlagen. Lediglich 600 Kombinationen von Runs und Topics in Deutsch sowie 900 Kombinationen von Runs und Topics in Englisch lagen vor. Daher wurde ein System für maschinelles

Lernen eingesetzt, das diese Daten sowohl auf ein lineares als auch nicht-lineares Modell überprüfte.

Als Lernsoftware kam das Waikato Environment for Knowledge Analysis (WEKA[5]) zum Einsatz, das zahlreiche Lernalgorithmen implementiert. WEKA wurde mit JAVA programmiert und steht als Open-Source Software zur Verfügung (Witten & Frank 2000).

Lernziel war die Retrieval-Qualität. Dabei wurde neben der absoluten Qualität auch die Abweichung vom Durchschnitt eingesetzt als auch die Qualitäts-Gruppen. Das Lernmodell sollte also von den Eigenschaften des Runs und des einzelnen Topic auf die Qualität dieses Runs bei dem Topic schließen.

Weder für die deutschen noch für die englischen Topics konnten Modelle für die gewünschte Abbildung gefunden werden. Dies galt für lineare Naive Bayes Algorithmen ebenso wie für nicht-lineare neuronale Netze.

5 Zusammenfassung

Die vorliegende Untersuchung konnte keine Einflüsse der Topic-Eigenschaften auf die Qualität der Ergebnisse der teilnehmenden Systeme feststellen. Demnach verfälschen die Topics die Ergebnisse nicht. Aus der Perspektive der CLEF Organisatoren zeigt dies, dass die Validität der Ergebnisse nicht durch die Formulierung der Topics beeinflusst wird. Aus Perspektive der System-Entwickler gibt es momentan keine Ansätze für die Verbesserung von Systemen durch die Analyse von Topics.

Die Untersuchung soll auf die CLEF 2002 Ergebnisse und weitere Topic Sprachen ausgedehnt werden. Zudem werden weitere automatisch erkennbare sprachliche Eigenschaften der Topics analysiert, wie etwa Satz-Komplexität oder Part-of-Speech Statistiken (cf. Allan & Raghavan 2002).

Die Analyse von Anfrage-Eigenschaften kann bei anderen als textuellen Objekten des Information Retrieval als wesentlich bedeutender eingeschätzt werden. Die Heterogenität der Anfragen steigt bei Audio-, Bild- und Video-Retrieval noch an (cf. Smeaton et al. 2002).

[5] http://www.cs.waikato.ac.nz/~ml/weka/index.html

Literatur

Allan, James/Raghavan, Hema (2002): Using part-of-speech Patterns to Reduce Query Ambiguity. In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ´02) Tampere. 307-314.

Baeza-Yates, Ricardo/Ribeiro-Neto, Berthier (Hrg.) (1999): Modern Information Retrieval. Harlow et al.: Addison-Wesley.

Evans, David/Grefenstette, Gregory/Qu, Yan/Gent, Joop van (2002): Anatomy of a CLIR Application. Invited Talk at the CLEF Workshop 2002. 20. Sept. Rom. http://clef.iei.pi.cnr.it:2002/workshop2002/presentations/com-clir.pdf

Kando, Noriko/Aihara, K/Eguchi, K./Kato, H. (Hrg.) (2001): Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, National Institute of Informatics (NII).

Kluck, Michael/Mandl, Thomas/Womser-Hacker, Christa (2002): Cross-Language Evaluation Forum (CLEF): Europäische Initiative zur Bewertung sprachübergreifender Retrievalverfahren. *nfd Information – Wissenschaft und Praxis* 53 (2), 82-89.

Kluck, Michael/Womser-Hacker, Christa (2002): Inside the Evaluation Process of the Cross-Language Evaluation Forum (CLEF): Issues of Multilingual Topic Creation and Multilingual Relevance Assessment. In: 3rd Intl. Conf. on Language Resources and Evaluation, Las Palmas, Spain.

Kwok, K. L./Chan, M. (1998): Improving Two-Stage Ad-Hoc Retrieval for Short Queries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ´98) Melbourne, 250-256.

Mandl, Thomas (2001): Tolerantes Information Retrieval: Neuronale Netze zur Erhöhung der Adaptivität und Flexibilität bei der Informationssuche. Konstanz: Universitätsverlag.

Mandl, Thomas/Womser-Hacker, Christa (2000): Ein adaptives Information Retrieval Modell für Digitale Bibliotheken. In: Knorz, Gerhard; Kuhlen, Rainer (Hrsg.): Informationskompetenz - Basiskompetenz in der Informationsgesellschaft. Proceedings 7. Intl. Symposium für Informationswissenschaft. (ISI 2000). Konstanz: Universitätsverlag. 1-16.

Peters, Carol (Hrg.) (2001): Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop. Berlin et al.: Springer (=LNCS 2069).

Peters, Carol/Braschler, Martin/Gonzalo, Julio/Kluck, Michael (Hrg.) (2002): Evaluation of Cross-Language Information Retrieval Systems. Proceedings of the CLEF 2001 Workshop. Berlin et al.: Springer (=LNCS 2406).

Savoy, Jacques (2002): Report on CLEF-2002 Experiments: Combining Multiple Sources of Evidence. In: Peters, Carol (Hrg.): Cross Language Evaluation Forum: Results of the CLEF 2002 Evaluation Campaign. Working Notes CLEF-Workshop. 19.-20.9.2002. Rom. 31-46.

Smeaton, Alan; Over, Paul; Taban, Ramazan (2001): The TREC-2001 Video Track Report. In: Voorhees & Harman 2001.

Sparck Jones, Karen (1995): Reflections on TREC. *Information Processing & Management* 31, 291-314.

Tague-Sutcliffe, J./Blustein, J. (1993): A Statistical Analysis of the TREC-3 Data. In: Harman, Donna (Hrsg.): The Third Text REtrieval Conference (TREC-3). NIST Special Publication. National Institute of Standards and Technology. Gaithersburg, Maryland, 385.

Oard, Douglas (1997): Serving Users in Many Languages: Cross-Language Information Retrieval for Digital Libraries. In: D-Lib Magazine, December 1997. http://www.dlib.org/dlib/december97/oard/12oard.html

Voorhees, Ellen (1998): Variations in relevance judgments and the measurement of retrieval effectiveness. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ´98) Melbourne, 315-223.

Voorhees, Ellen/Harman, Donna (Hrg.) (2001): The Tenth Text Retrieval Conference (TREC-10). NIST Special Publication. National Institute of Standards and Technology. Gaithersburg, Maryland. http://trec.nist.gov/pubs/

Wilkinson, Ross; Zobel, Justin; Sacks-Davis, Ron (1995): Similarity Measures for Short Queries. In: Harman, Donna (Hrg.): The Fourth Text REtrieval Conference (TREC-4). NIST Special Publication. National Institute of Standards and Technology. Gaithersburg, Maryland. http://trec.nist.gov/pubs/

Witten, Ian; Frank, Eibe (2000): Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations.

Womser-Hacker, Christa (1997): Das MIMOR-Modell. Mehrfachindexierung zur dynamischen Methoden-Objekt-Relationierung im Information Retrieval. Habilitationsschrift. Universität Regensburg, Informationswissenschaft.

– (2002): Multilingual Topic Generation within the CLEF 2001 Experiments. In: Peters et al. (2002) 389-393.