Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

Information Engineering 1: Information Retrieval

Kapitel 7 Advanced Topics: Mehrsprachigkeit und Sprachübergreifende Suche

M. Braschler

(basierend u.a. auf Material von J. Savoy)

"Themenkarte"

- Motivation
- „Jenseits von Englisch“: monolinguales IR
- Sprachübergreifendes IR (CLIR)
- Übersetzungsprobleme
- Übersetzungsstrategien
- Multilinguales IR

Lernziel Kapitel

- Verstehen, welche Unterschiede für verschiedene Sprache bestehen, und wie diese IR beeinflussen
- IR-Systeme für verschiedene Sprachen anpassen können
- Ansätze für sprachübergreifende Suche kennen

The challenge

"Given a query in any medium and any language, select relevant items from a multilingual multimedia collection which can be in any medium and any language, and present them in the style or order most likely to be useful to the querier, with identical or near identical objects in different media or languages appropriately identified."

[D. Oard & D. Hull,
AAAI Symposium on
Cross-Language IR,
Spring 1997, Stanford]

Motivation für “IR jenseits von Englisch”

6,800 “lebendige Sprachen” weltweit,

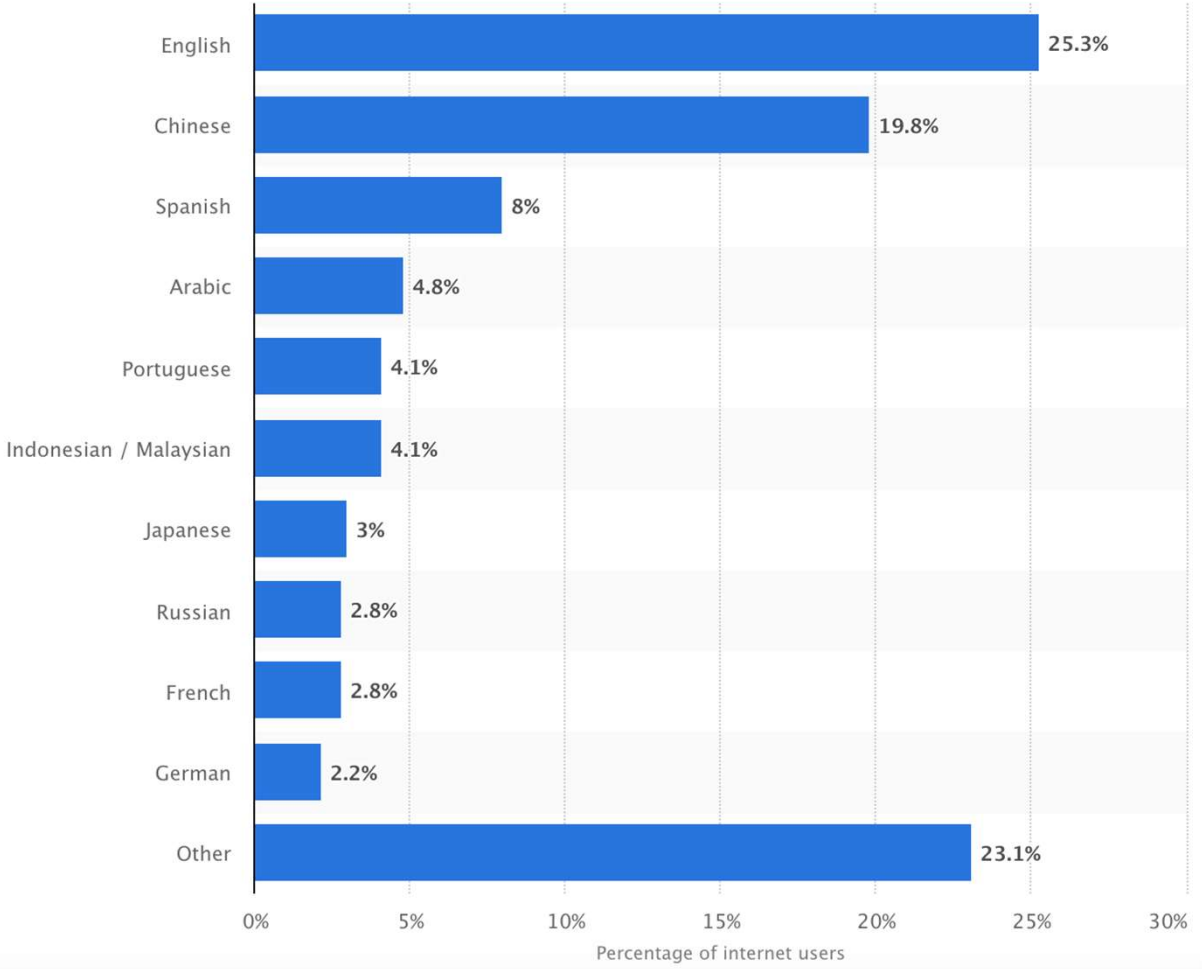
- 2,197 in Asien
- 2,092 in Afrika
- 1,310 im pazifischen Raum
- 1,002 in Amerika
- 230 in Europa.

Aber davon werden nur 600 geschrieben

- 80% der Weltbevölkerung spricht eine oder mehrere von 75 Sprachen
- 40% der Weltbevölkerung entfällt auf 8 verschiedene Sprachen
  - 75 Sprachen haben mehr als 10 Mio. Sprecher/innen
  - 20 Sprachen haben mehr als 50 Mio. Sprecher/innen
  - 8 Sprachen haben mehr als 100 Mio. Sprecher/innen

Siehe www.ethnologue.com und www.omniglot.com

Internet-User im Jahr 2020



Quelle: Statista

Sprache ist komplex

Eine Sprache hat

- Grössenordnung 100,000 Wörter
- Grössenordnung 10,000 syntaktische Regeln
- Grössenordnung 1,000,000 semantische Elemente

→ Auch für Menschen komplex zu lernen

→ Aber: in welchem Ausmass müssen wir sprachspezifische Eigenschaften für die ***Suche*** überhaupt berücksichtigen?

Motivation für sprachübergreifendes IR

- Bilingualität/Multilingualität (europa.eu)
- Viele Staaten oder Staatengebilde sind bi-/multilingual (Kanada (2), Singapur (2), Indien (21), EU (23))
- Offizielle Sprachen der EU: BG, CS, DA, DE, EL, EN, ES, ET, FI, FR, GA, HU, IT, LT, LV, MT, NL, PL, PT, RO, SK, SL und SV.
  Weitere Sprachen z.B.: Katalanisch, Galizisch, Baskisch, Walisisch, …
- Arbeitssprachen in der EU (hauptsächlich): Englisch, Deutsch, Französisch;
- In der UNO: Arabisch, Chinesisch, Englisch, Französisch, Russisch, Spanisch.
- Gerichtsentscheide werden in verschiedenen Sprachen verfasst
- Internationale Organisationen und Unternehmen: FIFA, WTO, Nestlé, …

Warum Resultate, die niemand lesen kann?

- Mehrsprachige Personen drücken ein Inforationsbedürfnis in einer Sprache aus, können aber verschiedene Sprachen lesen
- Passives Sprachverständnis: kann nicht die Frage in der gewünschten Sprache ausdrücken, versteht aber die Antwort
- “Entziffern” von sehr kurzen oder nicht-textuellen Antworten (Fakten, Statistiken, Bilder, Musik, …)
- Abklärung des Vorhandenseins von Treffern (und damit Vorlage zur Übersetzung)

Die Herausforderung

<TOPIC>

<TITLE>時代華納，美國線上，合併案，後續影響</TITLE>

<DESC> 查詢時代華納與美國線上合併案的後續影響。</DESC>

<NARR>

<BACK>時代華納與美國線上於2000年1月10日宣佈合併，總市值估計為3500億美元，為當時美國最大宗合併案。</BACK>

<REL>評論時代華納與美國線上的合併對於網路與娛樂媒體事業產生的影響為相關。敘述時代華納與美國線上合併案的發展過程為部分相關。內容僅提及合併的金額與股權結構轉換則為不相關。</REL>

</NARR>

<CONC>時代華納，美國線上，李文，Gerald Levin，合併案，合併及採購，媒體業，娛樂事業</CONC>

</TOPIC>

Die Herausforderung

- Other examples
- Strč prst skrz krk
- Mitä sinä teet?
- Mam swoją książkę
- Nem fáj a fogad?
- Er du ikke en riktig nordmann?
- Добре дошли в България!
- Fortuna caeca est
- نهارسعيد

Internationalisierung in Evaluationskampagnen

- Getrieben durch CLEF (ehemals: «Cross-Language Evaluation Forum») in den 00er-Jahren
  - Gestartet in 2000 mit EN, DE, FR, IT
  - 2001-02: EN, DE, FR, IT, SP, NL, FI, SW
  - 2003: DE, FR, IT, SP, SW, FI, RU, NL
  - 2004: EN, FR, RU, PT
  - 2005-06: FR, PT, HU, BG
  - 2007: HU, BG, CZ
  - 2008-09: Persisch
  - Monolingual, bilingual und multilingual
- Seit 2010: weitere Tasks (“Labs”): Image, Video, Fake News, Object Identification, Plagiatsdetektion, ….
- Weitere Kampagnen: NTCIR (JA, KR, ZH), TREC, FIRE (Hindi, Bengali, Marathi)…

Mehrsprachige Topics

- Topics (Informationsbedürfnisse) sind verfügbar in verschiedenen Sprachen (CLEF)
  - EN: Nestlé Brands
    FR: Les Produits Nestlé
    PT: Marcas da Nestlé
    HU: Nestlé márkák
    BG: Продуктите на Нестле
  - EN: Italian paintings
    FR: Les Peintures Italiennes
    PT: Pinturas italianas
    HU: Olasz (itáliai) festmények
    BG: Италиански картини

Monolinguales IR «jenseits von Englisch»

?

Welche sprachspezifischen Probleme sollten wir berücksichtigen, wenn wir uns «jenseits von Englisch» bewegen?

Monolingual IR: Schreibsysteme

- Verschiedene Alphabete
  - Lateinisch (26)
  - Kyrillisch (33)
  - Arabisch (28)
  - Hebräisch
  - Andere asiatische Sprachen: Hindi, Thai
- Silbenschriften
  - Japanisch: Hiragana (46) における
  - Katakana (46) フランス
  - Koreanisch: Hangul (8,200) 정보검색시스템
- Ideogramme
  - China (13,000/7,700) 中国人, Japan (8,800) ボ紛争
- Eine Transliterierung/Romanisierung ist (manchmal) möglich, siehe LOC - www.loc.gov/catdir/cpso/roman.html

Monolingual IR: Codepages

- Texte können mit verschiedenen Codepages dargestellt werden
- Das eigentliche ASCII-Encoding ist auf 7 Bit limitiert
- Für weitere Zeichen: Windows, Macintosh, BIG5, GB, EUC-JP, EUC-KR, …
- ISO-Latin-1 (ISO 8859-1 West European), Latin-2 (East European), Latin-3 (South European), Latin-4 (North European), Cyrillic (ISO-8859-5), Arabic (ISO-8859-6), Greek (ISO-8859-7), Hebrew (ISO-8859-8), …
- Unicode (UTF-8, see www.unicode.org)

Monolingual IR: Tokenisierung/Segmentierung

- Was ist ein Wort? Ein Token? Was sind “Zeichen”?

  I'll send you Luca's book

  C|net & Micro$oft

  IBM360, IBM-360, ibm 360, …

  Richard Brown

  brown paint

  Brown is the …

  Database system

  data base system

  data-base system

Monolingual IR: Tokenisierung/Segmentierung

- Wo startet/endet ein Wort?
- Zusammengesetzte Wörter finden sich in vielen Sprachen (siehe auch Englisch: worldwide, handgun, …). In manchen Sprachen ist dieser Wortbildungsprozess häufig (DE, NL, FI, HU, BG)
- In DE: "Bundesbankpräsident" =
  "Bund" + es + "Bank" + "Präsident"
  federal bank CEO
- Im Deutschen zentral: "Computersicherheit" – kann auch phrasal umschrieben werden, z.B. "die Sicherheit von Computern"
- Automatisches Decompounding steigert Retrievaleffektivität (+23% MAP, kurze Anfragen, +11% lange Anfragen, [Braschler & Ripplinger 2004].

Monolingual IR: Tokenisierung/Segmentierung

Ein schwieriges Problem in ostasiatischen Sprachen (JA, KR, ZH, …)

- Verschiedene Strategien sind denkbar («longest match», «mutual information», «dynamic programming», morphologische Analyse, …)

我不是中国人

我 不 是 中国人

I not be Chinese

Monolingual IR: Tokenisierung/Segmentierung

Im Japanischen Schreibsystemwechsel nutzen

コソボ紛争におけるNATOの攻撃と

Kanji (Chinesische ideograms)	42.3 %
Hiragana (e.g., in, of, …)	32.1 %
Katakana (e.g., フランス)	7.9 %
Romaji (our alphabet)	7.6 %
…andere	10.1 %

→ Chasen morphological analyzer (chasen.aist-nara.ac.jp)

Monolingual IR: Tokenisierung/Segmentierung

Dasselbe Konzept kann durch unterschiedliche Segmentierungen dargestellt werden

정보 (information) 검색 (retrieval) 시스템 (system)

정보검색 (information retrieval) 시스템 (system)

정보 (information) 검색시스템 (retrieval system)

정보검색시스템

→ Hangul Analyser Module (nlp.kookmin.ac.kr)

Monolinguales IR: sprachunabhängig

- Ein sprachunabhängiger Ansatz ist möglich: n-Gram-Indexierung
- Für die Grundlagen, siehe [McNamee & Mayfield 2004]
- Automatische Segmentierung von Sätzen (über Wortgrenzen hinweg)
- Verschiedene Spielarten möglich
- “The White House”
  → “The “, “he W”, “h Wh”, “ Whi”, “Whit”, “hite”, …
  oder
  → “the“, “whit”, “hite”, “hous”, “ouse”
- Ist oft ein effektiver Ansatz, wenn wenig Ressourcen für eine neue oder wenig gesprochene Sprache zur Verfügung stehen.
- Stoppwortelimination ist wahrscheinlich ebenfalls eine gute Idee, aber statistisch machbar (Wörter zählen)
- Klassische Strategie für JA, ZH und KR

?

Welche Implikationen hat so ein n-Gram-Ansatz für die Indexgrösse?

Was sind die Implikationen für die Ausbeute? Für die Präzision?

Monolinguales IR: Segmentierung

Verschiedene Darstellungsformen eines Satzes in Chinesisch:

我不是中国人

Unigrams
我 不 是 中 国 人
Bigrams
我不 不是 是中 中国 国人
Unigrams and bigrams
我, 不, 是, 中, 国, 人, 我不, 不是, 是中, 中国, 国人

Words (MTSeg)
我 不 是 中国人

Monolinguales IR: Segmentierung

Verschiedene Darstellungsformen eines Satzes in Japanisch:

クロソフトのWindowsがどのような競合関係

Unigrams

クロソフトWindows競合関係

Bigrams

クロ ロソ ソフ フト Windows 競合 合関 関係

Unigrams and bigrams

クロソフトWindows競合関係 クロ ロソ ソフ フト

競合 合関 関係

Words (ChaSen)

クロソフト Windows 競合 関係

Monolinguales IR: Segmentierung

ZH: Unigram & bigram > word (MTool) ≈ bigram

Eine echte Wortsegmentierung ist schwierig, resp. nicht wirklich effektiver (siehe auch z.B. [Abdou & Savoy 2006])

JA: Unigram & bigram ≈ word (Chasen) ≥ bigram

KR: bigram ≈ HAM (decompounding) > unigram

Monolinguales IR: Diakritische Zeichen

- Diakritische Zeichen
  - Unterscheiden sich von Sprache zu Sprache (“résumé”, “Äpfel”, “leão”)
  - Modifizieren die Bedeutung (z.B., “tache” (Fleck) or “tâche” (Aufgabe))
  - Manchmal aber auch verwandt (z.B., “cure” and “curé” Pfarramt / Priester
  - Werden meistens durch das IR-System entfernt
    - Unterschied normalerweise gering und nicht statistisch signifikant
    - Werden in vielen Sprachen nicht konsistent eingesetzt (Deutsch: alternative Schreibweise wenn nicht auf Tastatur, Französisch: weggelassen bei Grossschreibung)
- In manchen Sprachen gibt es auch Probleme mit inkonsistenter Rechtschreibung (color/colour, Stängel/Stengel, …)

Monolinguales IR: Namen

- Normalisierung / Namen
  - Homophone Namen: z.B. Stephenson (Dampfmaschine) und Stevenson (Autor) werden in Japanisch, Chinesisch, Koreanisch gleich ausgesprochen resp. geschrieben.
  - Die Schreibung von Namen differiert von Sprache zu Sprache (Gorbatschow, Gorbachev, Gorbacheff, Gorbachov)
  - Übersetzte Namen: Mona Lisa → La Joconde → La Gioconda
  - Hier sind spezialisierte Thesauri und Namenslisten sinnvoll

Monolinguales IR: Stoppwortlisten

- Stoppwortlisten
  - Sehr häufige, inhaltsarme Terme (Artikel, Präpositionen, Partikel, Konjunktionen, …
  - Problematisch im multilingualen Kontext (“or” – “oder” in Englisch, “Gold”/”nun” in Französisch, "who“ – “wer” in Englisch oder WHO (“World Health Organization”)
  - Ist auch systemabhängig (Fragewörter sind evtl. wichtig für QA-Systeme)
  - Könnte auch anfragespezifisch sein (nur die Wörter entfernen, die häufig in Anfragen vorkommen)

Monolinguales IR: Stemming

- Stemming (regelbasiert)
- Inflektionen
  Zahl (Singular / Plural): horse/horses, Tanne/Tannen
  Geschlecht: actress/actor, Jurist/Juristin
  Verbformen: jump/jumping/jumped, gehen/gehst/ging/gegangen
  Eher einfach im Englischen (‘-s’, ‘-ing’, ‘-ed’)
- Derivationen (Stamm + Suffix = Wort)
  ergibt neue Wörter (andere Wortart)
  ‘-ably’, ‘-ment’, ‘-ship’, …
  admit → {admission, admittance, admittedly}
  ‘-lich’, ‘-keit’, ‘-ung’, …
  Glück → {glückhaft, glücklich, glücken}

Monolinguales IR: Stemming

- Verschiedene Ansätze existieren:
  - Kein Stemming (resp. n-Gram!)
  - Inflektionale Stemmer (S-stemmer)
  - Derivationale Stemmer (Porter, Lovins, SMART)
- Verwenden des "Lemma", Resultat einer morphologischen Analyse
  - Ggf. Wortart berücksichtigen («Part-of-Speech»)
  - Synset gemäss WordNet

Monolinguales IR: Stemming

- S-Stemmer funktionieren gut für romanische Sprachen (FR, PT, …)
- Französisch: (nach Savoy)
    - For words of six or more letters
      if final letters are ‘-aux’ then replace ‘-aux’ by ‘-al’,
      if final letter is ‘-x’ then remove ‘-x’,
      if final letter is ‘-s’ then remove ‘-s’,
      if final letter is ‘-r’ then remove ‘-r’,
      if final letter is ‘-e’ then remove ‘-e’,
      if final letter is ‘-é’ then remove ‘-é’,
      if final two letters are the same, remove the final letter

Monolinguales IR: Stemming

- In germanischen Sprachen (Deutsch!) komplexer
- Verschiedenste Formen für den Plural (+ inkl. Modifikationen mit diakritischen Zeichen)
  “Motor”, “Motoren”; “Jahr”, “Jahre”;
  “Apfel”, “Äpfel”; “Haus”, “Häuser”
- Fälle führen zu verschiedenen Endungen:
  (z.B. Genitiv: “Staates”, “Mannes”)
  gilt auch für die Adjektive
  (“einen guten Mann”)
- Kompositabildung
  (“Lebensversicherungsgesellschaftsangestellter”
  = life + insurance + company + employee)

Monolinguales IR: Stemming

- Noch komplexer in anderen Sprachen, wie der Finno-Ugrischen Sprachfamilie (Anzahl Fälle: 18 in HU, 15 FI)

ház	nominative (house)
házat	accusative singular
házakat	accusative plural
házzal	“with” (instrumental)
házon	“over” (superessive)
házamat	my + accusative sing.
házamait	my + accusative + plur.

- Im Finnischen kann sich der Stamm selbst ändern (z.B., “matto”, “maton”, “mattoja” (Teppich))
  → Tiefere morphologische Analyse ist nützlich
- + Kompositabildung
  (“internetfüggök”, “rakkauskirje”)

Monolinguales IR: Stemming

- Das Arabische hat nochmals andere Wortbildungsprozesse
- Stemming ist hier wichtig:
  Word = prefix + stem + pattern + suffix
- Stems sind drei/vier Zeichen

  ktb + CiCaC = **kitab**

kitab	a book
kitabi	my book
al**kitab**	the book
kitabuki	your book (femi)
kitabuka	your book (masc)
kataba	to write
katib	the writer (masc)
katibi	the writer (femi)
ma**ktab**	office
ma**ktaba**	library …

  - Schreibvarianten für Fremdwörter müssen berücksichtigt werden
  - Die Wurzel (“root”) ist zu aggressiv für IR

Monolinguales IR: Stemming

Durchschnittliche Verbesserung der Retrievaleffektivität (MAP)

- +4% Englisch
- +4% Niederländisch
- +7% Spanisch
- +9% Französisch
- +15% Italienisch
- +19% Deutsch
- +29% Schwedisch
- +34% Bulgarisch
- +40% Finnisch
- +44% Tschechisch

Sprachübergreifendes IR

- Als «sprachübergreifendes Retrieval» werden Szenarien bezeichnet, in welchen mit einer Anfrage in Sprache L1 (auch) Dokumente in anderen Sprachen als L1 gefunden werden sollen (L2…Lx)

Bilingual: L1 → L2

Multilingual: L1 → (L1), L2, … Lx

Es muss jeweils ein «Language Gap» überwunden werden

«Bridging the Language Gap»: Übersetzung

Ist eine schwierige Aufgabe, auch für menschliche Übersetzer/innen

- Rom, Italien
  "Please dial 7 to retrieve your auto from the garbage"
- Indien
  "Children soup"
- Kairo, Ägypten
  "Unaccompanied ladies not admitted unless with husband or similar"
- Auf einer japanischen Medikamentenverpackung:
  "Adults: 1 tablet 3 times a day until passing away"

C. Crocker: *Løst in Tränšlatioπ. Misadventures in English Abroad.* O'Mara Books, London, 2006

Übersetzungsprobleme

- Wort-für-Wort Übersetzung passt oft nicht:
  “horse” – “Pferd”?

  Ja: “horse-race” – “Pferderennen”

  Nein: “horse-fly” – “Breme”

  Nein: “horse sense” – gesunder Menschenverstand
- Stehende Ausdrücke
  - “eat a little crow” → “eine kleine Krähe essen”???
- Kognaten/falsche Freunde: “sensible” <> “sensibel”

  “Requests of Quebec” = “Demandes du Québec”
  “Demands of Quebec” = “Exigences posées par le Québec”
- → Die Übersetzung muss die Bedeutung erhalten, nicht wörtlich sein

BZ DIE STIMME BERLINS

Sacramento – Kaliforniens Gouverneur Arnold Schwarzenegger, 56, zeigt sich in Sacramento im „Detroit Pistons“-Trikot (Foto). Er hält eine Zeitung hoch, die über den Sieg der „Pistons“ in der NBA-Liga über die „Los Angeles Lakers“ berichtet und isst eine „Spezialität“ aus Michigan: Eine kleine Krähe, dazu Kartoffelkuchen mit Fleisch und Gemüse. Grund: Mit dem Sieg der Pistons über die Lakers im Finale der NBA verlor er eine Wette gegen Gouverneurskollegin Jennifer Granholm aus Michigan.

Machine Translation: damals

- “Tainted-Blood Trial”
  - Manually “L'affaire du sang contaminé”
  - Systran “Épreuve De Corrompu - Sang”
  - Babylon “entacher sang procès”
- “Death of Kim Il Sung”
  - Manually “Mort de Kim Il Sung”
  - Systran “La mort de Kim Il chantée”
  - Babylon “mort de Kim Il chanter”
  - Babylon “Tod von Kim Ilinium singen ”
- “Who won the Tour de France in 1995?”
  - Manually “Qui a gagné le tour de France en 1995”
  - Systran “Organisation Mondiale de la Santé, le, France 1995 ”

Machine Translation: heute

Google Translate 2021

- «Tainted-Blood Trial»: Befleckte Blutprobe
- «Death of Kim Il Sung»: Tod von Kim Il Sung
- «Who won the Tour de France in 1995?»: Wer hat 1995 die Tour de France gewonnen?

- ➔ Fortschritte für «ganze», grammatikalische Sätze, immer noch Probleme mit Stichwortanfragen

?

Ist sprachübergreifende Suche ein schwierigeres oder leichteres Problem als maschinelle Übersetzung?

Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

!

Sprachübergreifende Suche <> Übersetzung + monolinguale Suche

- Die Übersetzung ist nicht Selbstzweck
- Wir «übersetzen», um zu suchen/finden
- Der Übersetzungsprozess kann vor dem Benutzer versteckt werden

«Pseudo-Übersetzung»

Eine bessere Übersetzung liefert nicht immer bessere Retrievaleffektivität

Translation	Query	AP
EN (original)	U.N./US Invasion of Haiti. Find documents on the invasion of Haiti by U.N./US soldiers.	
Reverso	Invasion der Vereinter Nationen Vereinigter Staaten Haitis. Finden Sie Dokumente auf der Invasion Haitis durch Vereinte Nationen Vereinigte Staaten Soldaten.	40.07
Free	U N UNS Invasion von Haiti. Fund dokumentiert auf der Invasion von Haiti durch U N UNS Soldaten	72.14

Automatische Übersetzung

- Automatische Übersetzung führt zu Mehrdeutigkeiten
  - Mehrere Übersetzungen für jedes Wort
  - Es interessant, Übersetzungswahrscheinlichkeiten einzubeziehen (aber wie?) (MT ist eine Blackbox)
  - Anfrageexpansion kann helfen
- Wichtige Übersetzungsressourcen
  - Zweisprachige, mehrsprachige Wörterbücher (Wortlisten)
  - Namenslisten
  - “Parallel corpora”
  - “Compatible/comparable corpora” (thematisch, zeitlich, kulturell)
  - MT-Systeme

Automatische Übersetzung

- Übersetzung mit wenig Kontext ist ein schwieriges Problem
  - Anfrage (“Query translation”)
    Die Anfrage kann eine Mischung aus Phrasen und Stichwörtern sein
    Beispiel: “car woman bag and man walking in a street”
    Auch für einen Menschen zum Teil kaum zu interpretieren (“plate orange” – Teller? Orange? Orangefarbiger Teller?)
  - Tabellenköpfe/-legenden
  - Bildbeschriftungen
  - Stichwortlisten zu Objekten (Metadaten)

- Es kommen vornehmlich statistische Verfahren zum Zug

Übersetzungsstrategien: was übersetzen?

- Wir ignorieren das Problem einfach
  - Der Text ist eine fehlerhaft geschriebene Version einer anderen Sprache (“near-cognates”) und mit einigen Korrekturregeln können wir das Matching ermöglichen (z.B., Cornell at TREC-6, Berkeley at NTCIR-5) – funktioniert nur für stark verwandte Sprachen
- Anfrageübersetzung
  - Weniger Rechenaufwand
- Dokumentübersetzung
  - Offline, benötigt redundanten Speicherplatz
- Anfrage- und Dokumentenübersetzung
  - Aufwändig, kann sehr effektiv sein
- Entwicklung der Retrievaleffektivität für die grössten Sprachen: von initial ~50% (TREC-6) auf bis zu 100% (heute)

Übersetzungsstrategien

- Wörterbücher - Machine-readable bilingual dictionaries (MRD)
  - Häufig mehr als eine Übersetzung (alle verwenden? Nur die erste? Wie gewichten?)
  - OOV-Problem (Namen)
  - Wir benötigen die Grundformen (lemmata)
- Maschinelle Übersetzung - Machine translation (MT)
  - Off-the-shelf verfügbar
  - Qualität variabel
  - Probleme mit fehlendem Kontext
- Statistische Übersetzungsmodelle
  - Verschiedene Vorschläge
  - Kann aus paralellen oder vergleichbaren Korpora lernen
  - Kann an Domäne angepasst werden (aber: Aufwand!)

Übersetzungsstrategien

- Ein beliebter Ansatz ist eine Erweiterung der Anfrage – vor oder nach dem Retrieval

- "Pre-translation expansion"
  könnte ein Problem sein für MT
- "Post-translation expansion"
  kann die Retrievaleffektivität steigern

Query_orig → Query_orig_exp → "Übersetzung" → Query_dest → Query_dest_exp

- ? Was ist die Idee hinter diesen Retrievalschritten?

Übersetzungsstrategien

Query in English	the economic and (commercial relations) between Mexico and Canada
After pre-translation expansion	economic (commercial relations) mexico canada mexico free-trade canada trade mexican salinas cuba pact economies barriers
After pre-translation expansion and translation to Spanish	[económico equitativo][comercio negocio tráfico industria] [narración relato relación][Méjico México] Canadá [Méjico México][convenio comercial][comercio negocio tráfico industria] zona cuba salinas *[equal economic][trade business traffic industry][narrative story relationship][Mexico Mexico] Canada [Mexico Mexico][trade agreement] [trade business traffic industry] area cuba salinas*
After pre-translation expansion, translation to Spanish and post-translation expansion, stemmed	canada (liber comerci) trat ottaw dosm (acured paralel) norteamer (est un)(tres pais) import eu (vit econom) comerci (centr econom) (barrer comerc)(increment subit) superpot rel acuerd negoci *Canada (free trade) treaty Ottawa 2000 (side agreement) north American (united states) (three countries) import eu (vital economy) trade (central economy)(trade barrier)(sudden increase) superpower relationship agreement business*

Beispiel aus Ballesteros and Croft (1997)

Übersetzungsstrategien

- Parallele Corpora [Shakery& Zhai, 2012]
  - Schwierig zu bekommen
  - Kulturelle, thematische und zeitliche Differenzen sind wichtig
  - Das Web ist eine Quelle, aber auch andere Quellen (z.B. Gesetzestexte, Parlamentsprotokolle, …)
- Comparable Corpora
  - Inhalte aus denselben Domänen, «Alignment» auf Dokumentenebene

Query_orig → SearchResult_orig → «alignment» → SearchResult_dest → Query_dest

Übersetzungsstrategien

Beispiele für Alignments auf Dokumentenebene

Title of German document	Title of French document
'New York Times': Erster 'Sternenkrieg'-Laser getestet (*'New York Times': First 'Star Wars' laser tested*)	Etats-Unis: premier essai grandeur réelle d'un puissant laser de l'IDS (*United States: first full-scale test of a powerful SDI laser*)
Goria zu Besuch in Malaysia (*Goria on visit in Malaysia*)	Le premier ministre italien Goria en visite en Malaisie (*Italian Prime Minister Goria on visit in Malaysia*)
Condor-Maschine bei Izmir abgestürzt: Mutmasslich 16 Tote (*Condor plane crashes near Izmir: presumably 16 deaths*)	Un avion ouest-allemand s'écrase près d'Izmir: 16 morts (*A Western German plane crashes near Izmir: 16 deaths*)

Zitiert nach Peters et al. (2012)

Übersetzungsstrategien

«Pseudo-Übersetzungen», aus den alignierten Dokumenten gewonnen

English → French: offer	French → German: incendie (*fire*)
4.7494 offre (*offer*)	0.6321 brand (*fire*)
4.5737 offert (*offered*)	0.5130 feuer (*fire*)
4.2255 comptant (*in cash*)	0.4223 feuerwehr (*fire department*)
4.0475 pret (*loan*)	0.4160 brandstiftung (*arson*)
3.8256 opa (*tender offer*)	0.3865 flammen (*flames*)
3.7980 faite (*done*)	0.3703 brandursache (*cause of fire*)
3.7215 prevoit (*calculates*)	0.3686 sachschaden (*material damage*)
3.5656 echec (*failure*)	0.3623 braende (*fires*)
3.5602 intention (*intention*)	0.2779 brannte (*burnt*)
3.5171 engage (*hire*)	0.2581 ausgebrochen (*broke out*)

Zitiert nach Peters et al. (2012)

?

Welche Probleme treten auf, falls mehrere Übersetzungskandidaten für ein Wort existieren?

Übersetzungsstrategien

«Strukturierte Queries» können helfen [Hedlund *et al.* 2004]

Eine Phrasenübersetzung kann helfen

Evaluationskampagnen (vor allem NTCIR) verwenden bewusst viele Namen in den Anfragen
→ diese müssen behandelt werden (z.B. Namenslisten)

English original topic formulation	What research is ongoing to reduce the effects of osteoporosis in existing patients as well as prevent the disease occurring in those unafflicted at this time?
English keyword query	osteoporosis prevent reduce research
Humanly translated Finnish query	osteoporoosi ehkäistä lieventää tutkimus
English query translated from Finnish by machine-readable dictionary; structured	#sum(osteoporosis #syn(prevent avert obviate obstruct hinder) #syn(alleviate mitigate reduce weaken abate relieve ease lighten) #syn (examination exploration inquest investigation report research scrutinity study))

Out of Vocabulary (OOV)

- Out-Of-Vocabulary
  - Wörterbücher haben einen beschränkten Umfang (dies gilt auch für die Wörterbücher, die MT-Systemen zugrunde liegen!)
  - Hauptproblem sind Namen (geographisch, Personen, Produkte, …)
  - Die “korrekte” Übersetzung kann unklar/nicht eindeutig sein (z.B. nach Chinesisch)
- Übersetzungen können Teilweise aus dem Web gelernt werden (über Kontext, Punktsetzung, ..) [Y. Zhang et al. TALIP]

Kulturelle Unterschiede

Dasselbe Konzept kann regionsabhängige Übersetzungen haben

z.B. “Mobile phone” in Französisch
« Natel » in der Schweiz
« Cellulaire » in Quebec
« Téléphone portable » in Frankreich
« Téléphone mobile » in Belgien

Siehe auch «Handy» - was nicht wirklich korrektes Englisch ist

Übersetzung: Pivot Language

Motivation:

- Bessere Übersetzungsressourcen können für gewisse Sprachpaare verfügbar sein (v.a. von und nach Englisch)
- Statt z.B. DE->ES gehen wir DE->EN->ES
- Löst ggf. auch die Probleme mit Kompositabildung im Deutschen
  Beispiel:
  “Robbenjagd” = “Robben”(seals) + “Jagd” (hunting)) wird korrekt übersetzt ins Englische (“Seal hunting”)
  aber z.B. nicht nach Französisch (“Robbenjagd” ist OOV).

“Lexical Triangulation”:

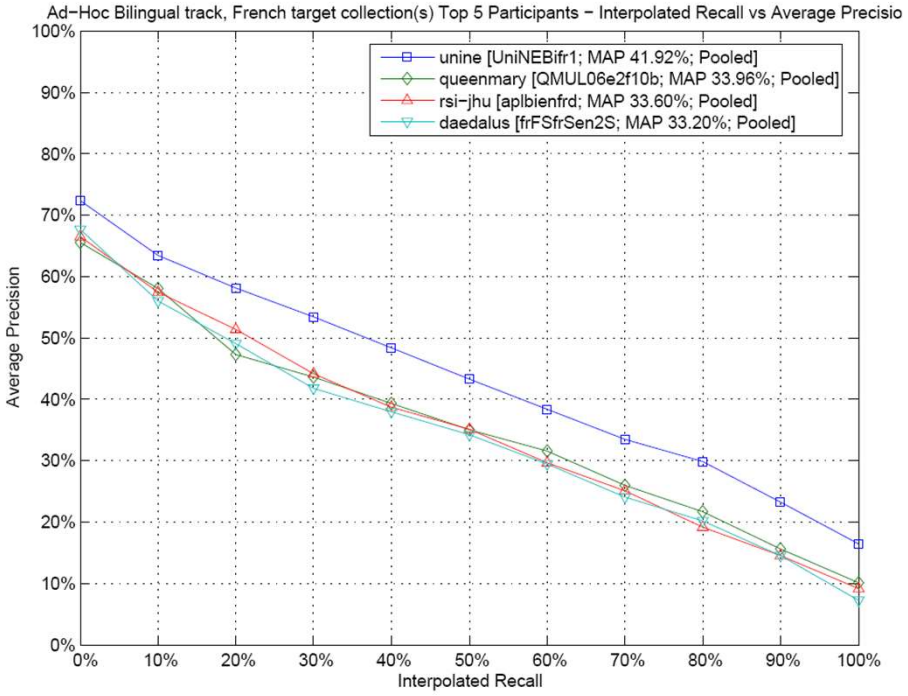
- Statt z.B. DE->ES wird DE->EN->ES und DE->FR->ES übersetzt, die Übersetzungen dann abgeglichen [Gollins & Sanderson]

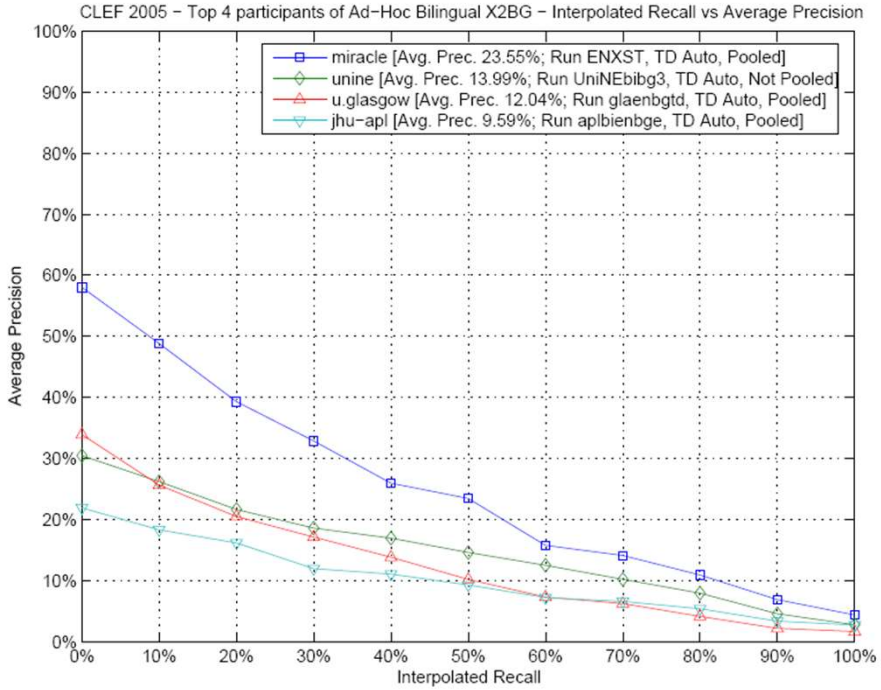
Bessere Übersetzung für IR: Ansätze

- Kontext nützen, Phrasenübersetzung
  - “temps” (FR) → time, weather, tense
  - “vol” (FR) → flight, theft, flock
  - “temps de vol” → time of flight
- Wortart nützen (EN → DE) (aber: wie in kurzer Anfrage bestimmen?)
  - “light” noun → “Licht”
  - Adjective → “hell”, “leicht”
- Domäne nützen (eindeutige Bedeutung = nur eine Übersetzung?)
- Window (Informatik) → OS?, windowing system, how to open a window in Java?, windows and UI?

Retrievaleffektivität

- Manche Sprachen sind «reifer» in Sachen sprachübergreifender Suche: verschiedene Ansätze funktionieren sehr gut, kleine Differenzen zu monolingual (minus 0-5%), Bsp. EN->FR, DE-FR etc.
- Andere Sprachen sind weniger gut erschlossen. Wenige gute Experimente, grosse Differenzen (-30% und mehr): Bsp. X->BG





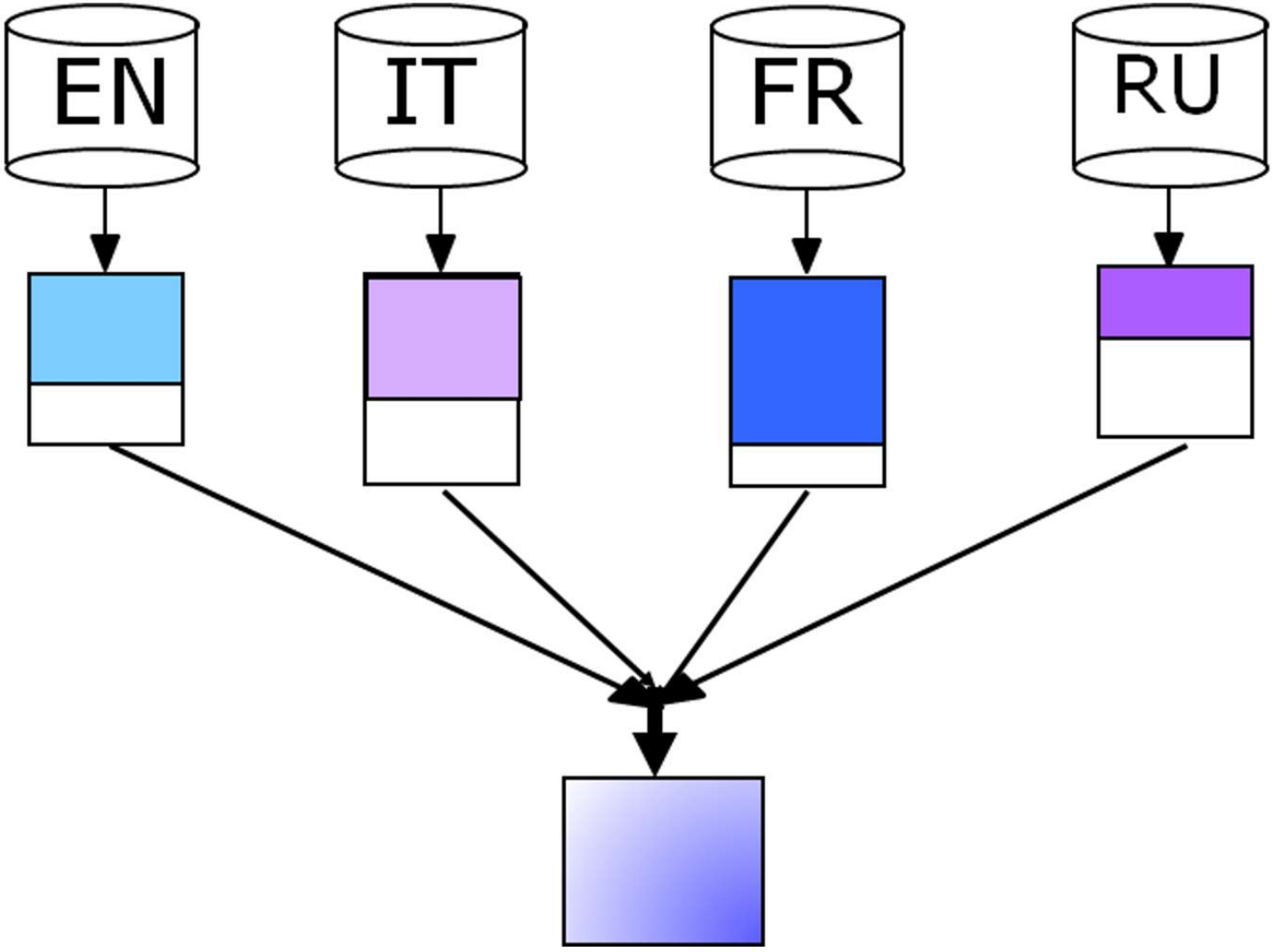
Mehrsprachiges IR: Multilingual IR

- Problem: Retrieval auf einer mehrsprachigen Kollektion
- Ansatz 1: multilingualer Index
  - Der Index enthält Dokumente in allen verschiedenen Sprachen
  - Die Anfrage wird in alle Sprachen übersetzt
  - Suche auf multilingualem Index → mehrsprachiges Resultat
  - Funktioniert schlecht (wieso?)
- Ansatz 2: ein gemeinsamer Index mit übersetzten Dokumenten (document translation (DT))
  - Alle Dokumente werden in eine gemeinsame Interlingua übersetzt (z.B. Englisch, oder abstrakte Darstellung)
  - Suche auf diesem Index, Resultat ist eine gemeinsame Rangliste
  - Effektiv, aber grosser Aufwand (Speicher/Rechenzeit)

Mehrsprachiges IR: Multilingual IR

- Ansatz 3: “Query translation (QT)” und Merging
  - Alle Anfragen in alle Dokumentensprachen übersetzen
  - Suche auf Einzelindexen für jede Sprache (parallel)
  - Ranglisten zusammenführen (Merging) (schwieriges Problem!)
  - Skaliert schlecht mit der Anzahl der Sprachen
- Ansatz 4: QT und DT mischen – grosser Aufwand, aber effektiv
- Ansatz 5: keine Übersetzung
  - Nur für stark verwandte Sprachen/Schreibsysteme
  - Limitierte Anwendbarkeit (Namen, auf Bildern, …)

Mehrsprachiges IR: Multilingual IR



Mehrsprachiges IR: Multilingual IR

Merging-Problem: siehe auch Websuche resp. verteiltes IR

1	EN120	1.2
2	EN200	1.0
3	EN050	0.7
4	EN705	0.6
...		

1	FR043	0.8
2	FR120	0.75
3	FR055	0.65
4	...	

1	RU050	6.6
2	RU005	6.1
3	RU120	3.9
4	...	

Mehrsprachiges IR: Multilingual IR

- Round-robin – immer im Turnus ein Dokument “ziehen”
- Raw-score merging – “rohe” Scores verwenden
  Spielart: wenn ein Dokument in mehreren Teilindexen/Teilresultaten vorkommen kann:

$$RSV(D_i) = \sum_{j=1}^{k} Score_j(D_i)$$

- Normalisierung (z.B. aufgrund des Topscores) (Problem?)

$$RSV(D_i) = \sum_{j=1}^{k} Score'_j(D_i)$$
$$with\ Score'_j(D_i) = \frac{Score_j(D_i)}{ScoreMax_j}$$

Mehrsprachiges IR: Multilingual IR

- Biased round-robin – “historischen Wert” (gelernt auf Trainingsdaten) der Teilkollektionen einbeziehen, potentiell mehrere Dokumente im Turnus “ziehen”, evtl. anfragespezifisch

  Z-score

  basierend auf Mittelwert und Standardabweichung

$$RSV(D_i) = \sum_{j=1}^{k} Score'_j(D_i)$$

$$with\ Score'_j(D_i) = \frac{(Score_j(D_i) - \mu_j) + \delta_j}{\sigma_j}$$

- Logistic regression [Le Calvé 2000], [Savoy 2004]

$$Score'_j(D_i) = \frac{1}{1 + e^{-[\alpha_j + \beta_{1j} \cdot ln(rank(D_i)) + \beta_{2j} \cdot RSV(D_i)]}}$$

und viele weitere mehr….

Schlussfolgerungen

- Viele der grundlegenden Konzepte (“Fundamentals”) die wir besprochen haben, sind weitgehend sprachunabhängig gültig
- Monolingual
  - Es ist relativ einfach, eine starke englische Baseline auf andere Sprachen, die verwandt sind, anzupassen (romanische und germanische Sprachen)
  - Aber: Kompositabildung in DE
  - Schwieriger, wenn die Sprache weniger verwandt ist (Finno-ugrisch, slawisch, …): teilweise ist eine sorgfältigere morphologische Analyse zwingend
  - Segmentierung ist ein Problem (ZH, JA, KR)
  - Nicht für alle Sprachen haben wir gute Ressourcen
  - Auch die Testkollektionen sind nicht alle gleich gut entwickelt (z.B. Probleme mit AR (TREC) und RU (CLEF))
  - Im Extremfall: komplett sprachunabhängiger Ansatz (n-Gram)

Schlussfolgerungen

- Bilingual / Multilingual
  - Ist nicht das Gleiche wie MT + monolingual!
  - Gewisse Sprachpaare haben viele, gut entwickelte Tools (z.B. EN, kann für Ansätze wie Pivot Language und Lexical Triangulation genutzt werden)
  - Weniger verbreitete Sprachen bieten mehr Probleme
  - Effektivität im Idealfall auf monolingualem Niveau
  - Merging ist ein schwieriges Problem