Query and Document Translation by Text Categorization

Julien Gobeill, Henning Müller, Patrick Ruch
julien.gobeill@sim.hcuge.ch

CLEF 2006, Alicante, September 22-23

© 2003 Hôpitaux Universitaires de Genève

CLIR vs. MMTQ

- Remark: **C**.L.I.R vs. Manual Multilingual Translation of Queries
  - French + English + German queries ≠ CLIR
- Translation: What ?
  - Queries: usual way (for ex: Rosemblat and al., CLEF 2003)
  - Documents: rare/expensive (LOGOS - Oard and Hackett 1998)

  → Performance depends on the translation strategy
  Context-dependent or not ?
  - **Both: Medical Subject Headings as interlingua** !
- Machine Translation: **CLIR Ratio** = 60%
- Thesaurus-driven Lexicon: 65-75% (Eichmann and al. 1998)
- Text categorization: 80% (Ruch 2004)
- Bilingual Parallel Corpora: > **90%** (Dumais and al. 1997)
- Text categorization + Machine Translation: > **90%** (Ruch 2004)

© 2003 Hôpitaux Universitaires de Genève

Query Translation by Machine Translation

ORIGINAL QUERY (CF): What are the effects of calcium on the physical properties of mucus from cystic fibrosis patients ?

FRENCH (Human): Quels sont les effets du calcium sur les propriétés physiques du mucus chez les patients atteints de mucoviscidose ?

SYSTRAN: Which are the effects of calcium on the properties *physiques* of mucus among patients *reached* of mucoviscidose ?

→ ~Grammatical Translation ***but...***

© 2003 Hôpitaux Universitaires de Genève

Some Issues of MT for IR...

- Babel Fish (Systran's system in AltaVista)

  alcoholic fatty liver → foie gras alcoolique

  Stéatose hépatique alcoolique

> **ExGallia - La Boutique des Francais du Monde**
> ... vos abonnements partout dans le monde chez vous ! **Foie-gras** truffé du Lot, confit de canard à l ... Apéritif 100% naturel obtenu par la fermentation **alcoolique** de 6 à 700 fleurs par bouteilles. A ...
> www.exgallia.com/produits-francais02.htm • Translate

→ CLIR Ratio = 60%

→ ***... But*** available on the shelf !

© 2003 Hôpitaux Universitaires de Genève

Bilingual parallel corpora (source UN)

81. Avec la mise en oeuvre du système intégré de gestion (SIG), grâce à l'analyse des traces électroniques, les possibilités de contrôle et de vérification seront plus étendues que jamais. Le SIG marque une étape importante dans l'uniformisation et la rationalisation de la pratique de la gestion dans tous les lieux d'affectation de l'Organisation. Pour la première fois, l'ONU va pouvoir disposer en temps voulu d'une information complète et récente sur ses ressources et leur emploi. Utilisé par d'autres programmes et organismes des Nations Unies, le SIG pourrait également être un facteur de transparence et de plus grande compatibilité de l'information d'un organisme à l'autre, ce qui conduirait à une harmonisation sur le plan administratif.

81. With the implementation of the Integrated Management Information System (IMIS), greater monitoring and audit capabilities will be available through electronic audit trails than ever before. IMIS is a major step in standardizing and rationalizing the management process in the Organization across duty stations. The Organization will be able, for the first time, to have access to timely, up-to-date and comprehensive information on its resources and their utilization. The use of IMIS by other programmes and organizations in the United Nations system could also promote greater transparency and compatibility of information across organizations, leading to standardization in administrative matters.

→ Linear projection methods to built transfer matrices, CLIR Ratio = 90%

→ Problem: overkill to develop these resources if not available !

© 2003 Hôpitaux Universitaires de Genève

Text categorization…

Functional Example

FRENCH (Human): Quels sont les effets du calcium sur les propriétés physiques du mucus chez les patients atteints de mucoviscidose ?

Top 1: mucoviscidose
Top 2: calcium + mucoviscidose
Top 3: calcium + physique + mucoviscidose
Top 4: calcium + physique + mucoviscidose + humanités
Top 5: calcium + physique + mucoviscidose + humanités + mucus
...
Top N: {...}
→ BoW: {calcium, humanities, physics, *cystic fibrosis*, mucus}

Specific needs: French Stemmer (Savoy), but cognates are frequent !

→ What is the best threshold value for N ?

© 2003 Hôpitaux Universitaires de Genève

Resources and Work Plan

- A multilingual controlled vocabulary
  - Medical Subject Headings (~20 000 concepts)
  - Thesaurus: 120 020 (UMLS) – 5000 (UML*F*)
- Collection
  - OHSUMED
  - Tuning: OHSUMED queries translated in French by an expert
- Development (if possible language-independent)
  - Translation: Use the categorizer and MeSH [as interlingua] for CLIR purposes

© 2003 Hôpitaux Universitaires de Genève

Example of MEDLINE Record

PMID: 11924965

Simple multiplex genotyping by surface-enhanced resonance Raman scattering.

Graham D, Mallinder BJ, Whitcombe D, Watson ND, Smith WE.

The accurate detection of DNA sequences is essential for a variety of post human genome projects including detection of specific gene variants for medical diagnostics and pharmacogenomics. A specific DNA sequence detection assay based on surface-enhanced resonance Raman scattering (SERRS) and an amplification refractory mutation system (ARMS) is reported. Initially, generation of PCR products was achieved by using specifically designed allele-specific SERRS active primers. Detection by SERRS of the PCR products confirmed the presence of the sequence tested for by the allele-specific oligonucleotides. This lead directly to the multiplex genotyping of human DNA samples for the deltaF508 mutational status of the cystic fibrosis transmembrane conductance regulator gene using SERRS active primers in an ARMS assay. Removal of the unincorporated primers allowed fast and accurate analysis in this system in a multiplex format without any separation of amplicons. The results indicate that SERRS can be used in modern genetic analysis and offers an opportunity for the development of novel assays. This is the first demonstration of the use of SERRS in multiplex genotyping and shows potential advantages over fluorescence as a detection technique with considerable promise for future development.

Major MeSH: Cystic Fibrosis*; DNA*; Genotype*; Polymerase Chain Reaction

Minor MeSH: HLA-DQ Antigens; Human; Reverse Transcriptase;
Sequence Analysis; Spectrum Analysis, Raman;
Support, Non-U.S. Gov't

© 2003 Hôpitaux Universitaires de Genève

ATC General Strategies

- *Empirical learning of text-concept associations* from a training set of texts and their associated concepts:
  - Reuters (Bayesian classifiers, Lewis 1992): 100 classes
  - Text categorization/filtering paradygm [Sebastiani: hundreds…]

  → Effective but Learning Conditions and Scalability…

- *Retrieval based on word-matching*, which attributes concepts to text based on lexical similarities:
  - Cross Language IR (SAPHIRE Int., Hersh et al. 1998)
  - Recent and rare
  - Hypothesis: sufficient for mapping queries/documents and MeSH terms

© 2003 Hôpitaux Universitaires de Genève

Basic Classifiers

- C1: FSA Pattern matcher + thesaurus [RegEx]
  $word_1 \ldots word_n \rightarrow word_1 \ldots \_^{[*,2]} \ldots word_n$
  $word_1 \ldots word_n \rightarrow word_1 \ldots [word_i]^* \ldots word_n$
  → Boolean scoring

- C2: Vector Space: Porter stems + TF*IDF weighting [VS]
  → Cosine distance/Similarity/Pivoted normalization

- C': UMLS Thesaural resources

- C'': Linguistically-motivated indexing units (NP)

© 2003 Hôpitaux Universitaires de Genève

Metrics: Recall and Precision

- Relevant retrieved
  How many terms are found for complete run ?

- Mean Reciprocal Rank (Maximize precision)
  Precision of the top-ranked category

- Mean Average precision
  Average Precision over 11 recall points

© 2003 Hôpitaux Universitaires de Genève

Query Translation

What are the effects of calcium on the physical properties of mucus from cystic fibrosis patients ?

Top 1: cystic fibrosis
Top 2: calcium + cystic fibrosis
Top 3: calcium + physics + cystic fibrosis
Top 4: calcium + physics + cystic fibrosis + humanities
Top 5: calcium + physics + cystic fibrosis + humanities + mucus
Top 10: cystic fibrosis + calcium + physics + humanities + humanism + mucus + health physics + human rights + calcium compounds + physical therapy

→ ? Automatic Text Categorization ?:
Fine-tuning on an OHSUMED Document mapping task (200-300 t.) then **short** queries (7-30 tokens)

© 2003 Hôpitaux Universitaires de Genève

Query Length as Parameter

What are the effects of calcium on cystic fibrosis patients ?

0840|cystic fibrosis transmembrane conductance regulator|
1001|humanism|
1001|humanities|
→ T=3 2501|**calcium**|
2724|fibrosis|
6070|**cystic fibrosis**|

What are the effects of calcium on the physical properties of mucus from cystic fibrosis patients ?

1001|humanities|
→ T=4 1807|**mucus**|
2501|**calcium**|
2724|fibrosis|
6070|**cystic fibrosis**|

Maximum ~ 3 for query mapping →, what about document ?

© 2003 Hôpitaux Universitaires de Genève

Results

- Medical ImageCLEF

  Top 3: MAP = 0.1913
  Top 5: MAP = 0.1967
  Top 8: MAP = 0.2255 [GE_8EN.treceval.eval]
  […]
  Top 20 ?

© 2003 Hôpitaux Universitaires de Genève

Conclusion

- Data-poor Text Categorization is effective for CLIR
- Mostly language independent

- Try with more than 8 terms !
- Recompute fusion with image features !

© 2003 Hôpitaux Universitaires de Genève

Thank you for your attention…

-EAGL Consortium: Swiss National Foundation
-Swiss-Prot Group: Anne-Lise Veuthey, Violaine Pillet

© 2003 Hôpitaux Universitaires de Genève