Multilingual Search

Shibamouli Lahiri

(shibamouli@cse.psu.edu)

Existing Work in CLIR

- **Query Translation**

- **Document Translation**

- **No translation** – cognate matching, transliteration

- **Interlingual Techniques** (LSI[5], decipherment)

Existing Work in CLIR (Contd.)

- **Query Stemming** – before/after translation

- **Translation Knowledge** – Bilingual Dictionaries (online, offline)[1], Parallel Corpora, the Web[6]

- **Translation Disambiguation** – word co-occurrence statistics, translation probability, POS tags, user and pseudo relevance feedback and user-assisted translation, structured queries[9], bi-directional translation, using web resources[11, 12]

- **Evaluation** – TREC CLIR track, NTCIR, TDT, CLEF

Existing Work in CLIR (Contd.)

- **Ranking** – vector space model, language model

- **Pivot Language (Triangulation) Approach**[8]

- **Named Entity Extraction and Translation**[7]

CyDAR – Basic Premises and Requirements

- Corpus is in Ancient Greek
- Query might be in English, Modern Greek or Ancient Greek

Issues

- Unit of Retrieval (Pages? Part of a page? Paragraphs?)

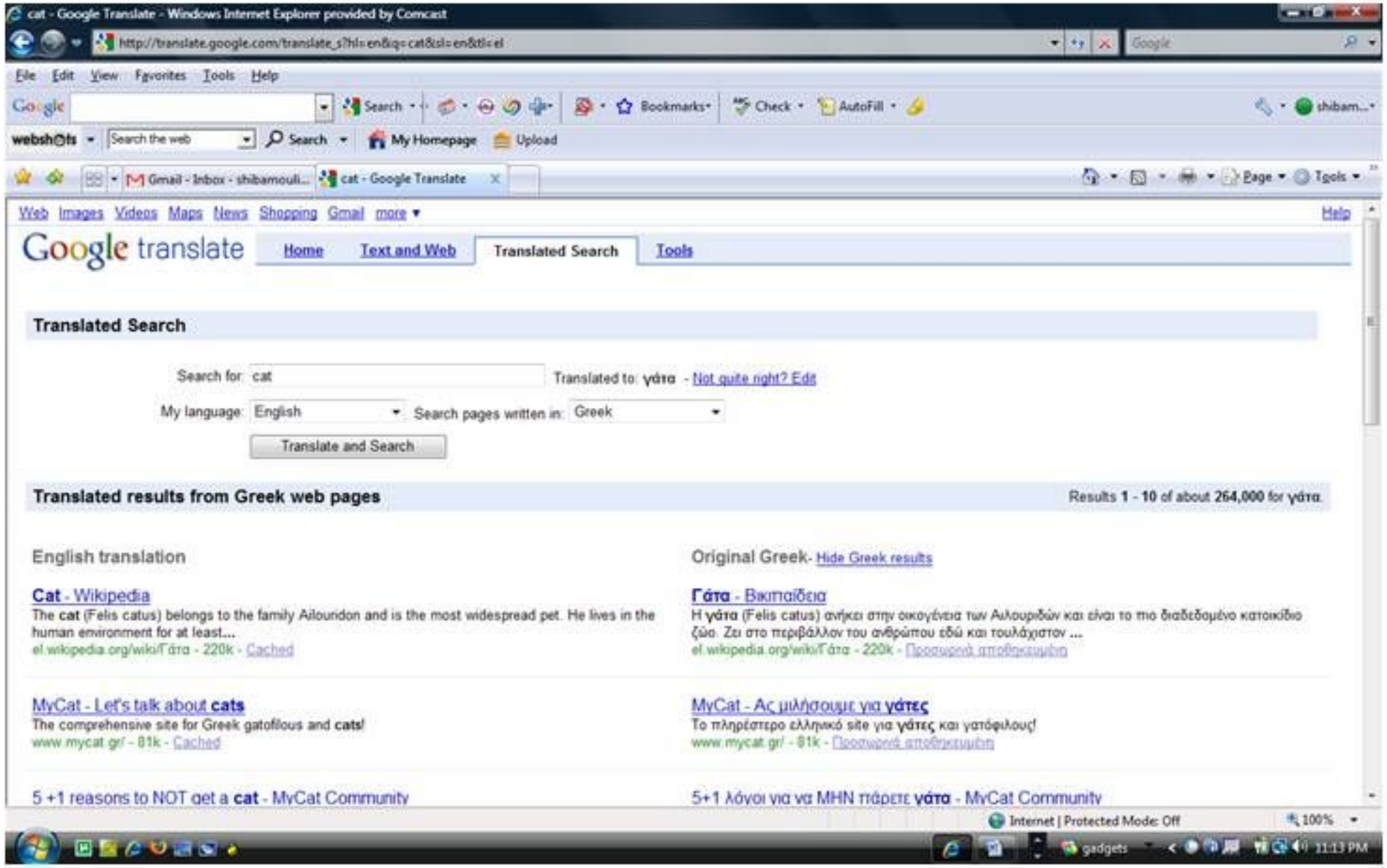
- An intelligible summary, preferably in all three languages

- Ranking

- Evaluation

- Query and translation disambiguation

User Interface Example



Lexicon

- Lexilogos

- Kypros (also offers Ancient Greek)

Parallel Corpora

- Perseus Project has English ←→ Ancient Greek and English ←→ Latin

- Europarl

- ELRA (European Language Resources Association)

References

1. **Technical issues of cross-language information retrieval: a review.** Kazuaki Kishida. *Information Processing and Management (2005).*
2. **Cross-Language Information Retrieval.** Daqing He and Jianqiang Wang. *Book chapter, Information Retrieval: Searching in the 21st Century (2009).* Wiley.
3. **Indian Language Information Retrieval.** Prasenjit Majumder and Mandar Mitra. *Book chapter, Guide to OCR for Indic Scripts.* Springer.
4. **Cross-Language Information Retrieval.** D.W. Oard and A.R. Diekema. *Annual Review of Information Science, Volume 33 (1998).*
5. **Automatic Cross-Language Information Retrieval using Latent Semantic Indexing.** M.L. Littman, S.T. Dumais and T.K. Landauer. *SIGIR Multilingual IR Workshop (1996).*
6. **Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web.** J. Nie, M. Simard, P. Isabelle and R. Durand. *ACM SIGIR Conference on Research and Development in Information Retrieval (1999).*

References (Contd.)

7. **Proper name translation in cross-language information retrieval.** H.H. Chen, S.J. Huang, Y.W. Ding and S.C. Tsai. *ACL (1998).*
8. **Improving cross language retrieval with triangulated translation.** T. Gollins and M. Sanderson. *ACM SIGIR Conference on Research and Development in Information Retrieval (2001).*
9. **Using Structured Queries for Disambiguation in Cross-Language Information Retrieval.** David A. Hull. *AAAI (1997).*
10. **Disambiguation Strategies for Cross-Language Information Retrieval.** D. Hiemstra and F. de Jong. *Lecture Notes in Computer Science (1999).* Springer.
11. **Using the Web for Translation Disambiguation.** Y. Zhang and P. Vines. *NTCIR-5 Workshop (2005).*
12. **Query Disambiguation for Cross-Language Information Retrieval Using Web Directories.** F. Kimura, A. Maeda, J. Miyazaki and S. Uemura. *Web Information Retrieval and Integration (2005).*

Thank you!