

# Multilingual Search

Shibamouli Lahiri

([shibamouli@cse.psu.edu](mailto:shibamouli@cse.psu.edu))

# Existing Work in CLIR

- **Query Translation**
- **Document Translation**
- **No translation** – cognate matching, transliteration
- **Interlingual Techniques (LSI<sup>[5]</sup>, decipherment)**

# Existing Work in CLIR (Contd.)

- **Query Stemming** – before/after translation
- **Translation Knowledge** – Bilingual Dictionaries (online, offline)<sup>[1]</sup>, Parallel Corpora, the Web<sup>[6]</sup>
- **Translation Disambiguation** – word co-occurrence statistics, translation probability, POS tags, user and pseudo relevance feedback and user-assisted translation, structured queries<sup>[9]</sup>, bi-directional translation, using web resources<sup>[11, 12]</sup>
- **Evaluation** – TREC CLIR track, NTCIR, TDT, CLEF

# Existing Work in CLIR (Contd.)

- **Ranking** – vector space model, language model
- **Pivot Language (Triangulation) Approach<sup>[8]</sup>**
- **Named Entity Extraction and Translation<sup>[7]</sup>**

# 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

The screenshot shows a Windows Internet Explorer window with the title bar "cat - Google Translate - Windows Internet Explorer provided by Comcast". The address bar contains the URL "http://translate.google.com/translate\_s?hl=en&iq=cat&sl=en&tgt=el". The menu bar includes File, Edit, View, Favorites, Tools, and Help. The toolbar has icons for Back, Forward, Stop, Refresh, Home, Bookmarks, Check, AutoFill, and a search bar with the placeholder "Search the web". Below the toolbar is a navigation bar with links to "My Homepage" and "Upload". The main content area is titled "Google translate" and shows the "Translated Search" tab selected. It displays a search form with "Search for: cat" and "Translated to: γάτα - Not quite right? Edit". The "My language: English" dropdown is set to English, and the "Search pages written in: Greek" dropdown is set to Greek. A "Translate and Search" button is present. The results section is titled "Translated results from Greek web pages" and shows two entries:

| English translation  | Original Greek - Hide Greek results   |
|--|---|
| <b>Cat - Wikipedia</b><br>The cat ( <i>Felis catus</i> ) belongs to the family Ailuridae and is the most widespread pet. He lives in the human environment for at least...<br><a href="http://el.wikipedia.org/wiki/%CE%93%CE%91%CE%94">el.wikipedia.org/wiki/%CE%93%CE%91%CE%94</a> - 220k - <a href="#">Cached</a> | <b>Γάτα - Βικιπαίδεια</b><br>Η γάτα ( <i>Felis catus</i> ) ανήκει στην οικογένεια των Αιλουρίδων και είναι το πιο διαδεδομένο κατοικίδιο ζώο. Ζει στο περιβάλλον του ανθρώπου εδώ και τουλάχιστον ...<br><a href="http://el.wikipedia.org/wiki/%CE%93%CE%91%CE%94">el.wikipedia.org/wiki/%CE%93%CE%91%CE%94</a> - 220k - <a href="#">Προσωρινή αποθήκευση</a> |
| <b>MyCat - Let's talk about cats</b><br>The comprehensive site for Greek catlovers and cats!<br><a href="http://www.mycat.gr/">www.mycat.gr/</a> - 81k - <a href="#">Cached</a>  | <b>MyCat - Άστεγοι για γάτες</b><br>Το πληρότερο ελληνικό site για γάτες και γατόφικους!<br><a href="http://www.mycat.gr/">www.mycat.gr/</a> - 81k - <a href="#">Προσωρινή αποθήκευση</a>   |

At the bottom, the status bar shows "Internet | Protected Mode: Off", "100%", and the time "11:13 PM".

# 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!**