Cross Lingual and Semantic Retrieval for Cultural Heritage Appreciation

Shuly Wintner
University of Haifa and Carnegie Mellon University
Prague, June 24, 2007

Idan Szpektor, Ido Dagan, Alon Lavie, Danny Shacahm and Shuly Wintner
Cross Lingual and Semantic Retrieval for Cultural Heritage Appreciation
ACL 2007 Workshop on Language Technology for Cultural Heritage Data
June 28, 2007, Prague, Czech Republic
A system for cultural heritage appreciation that employs state of the art human language technology

Semantic expansion and cross-lingual IR in the domain of archaeology and history of Israel

Specifically developed for the Hecht Museum in Haifa, Israel

Provides different users with different capabilities, bridging over language divides

Ultimately, it complements the visit to the museum with long-lasting instillation of information.
A domain-specific search engine that enables users to specify queries and retrieve information pertaining to the domain of the museum.

The engine is enriched by linguistic capabilities which embody an array of means for addressing semantic variation.

Queries are expanded using two main techniques:
- Semantic expansion based on textual entailment
- Cross-lingual expansion based on translation of Hebrew queries to English and vice versa

Retrieved documents are presented as links with associated snippets; snippets are translated from Hebrew to English.
The system was recently demonstrated successfully at the museum; it could be useful to a variety of museum visitor types, from children to experts.

More generally, it is an instance of adaptation of state of the art human language technology to the domain of cultural heritage appreciation, demonstrating how general resources and tools are adapted to a specific domain.

Finally, it provides a test-bed for evaluating the contribution of language technology in general, as well as specific components and resources, to a large-scale natural language processing system.
The Hecht Museum
The Hecht Museum
Museum visits are enriching experiences
The experience does not have to end when the visit ends
Further exploration of the artifacts and their influence on the visitor is possible after the visit
One common means of exploration is Information Retrieval
However, museum document collections are usually much smaller than general collections
The performance of human language technology, when applied to small collections, is hampered by two phenomena:

- **Ambiguity**: a text can be interpreted in more than one way
- **Semantic variability**: the ability to express a specific meaning in many different ways
  
  "Archaeologists found a new tomb"
  "Archaeologists discovered a tomb"
  "A sarcophagus was dug up by Egyptian Researchers"

- The same information can be expressed in different languages

- IR systems that do not address the variability in languages may suffer from lower recall, especially in restricted domains
Textual Entailment and Entailment Rules

- Inducing relations of semantic variability has been identified as a core semantic inference task by the generic textual entailment framework (Dagan, Glickman, and Magnini, 2006; Bar-Haim et al., 2006)
- The typical way to address variability in IR is to use lexical query expansion (Lytinen, Tomuro, and Repede, 2000; Zukerman and Raskutti, 2002)
- Another important type of knowledge representation is entailment rules and paraphrases
- An entailment rule is a directional relation between two templates, text patterns with variables, e.g., ‘X discover Y → X find Y’
- Paraphrases can be viewed as bidirectional entailment rules
The difficulties caused by variability are amplified when the user is not a native speaker of the language in which the retrieved texts are written.

While most Israelis can read English documents, fewer are comfortable with the specification of English queries.

In a museum setting, some visitors may be able to read Hebrew documents but still be poor at searching for them.

Other visitors may be unable to read Hebrew texts, but still benefit from non-textual information (e.g., pictures, maps, audio and video files, external links, etc.)
We developed a search engine that is capable of performing cross-lingual (Hebrew-English and English-Hebrew) information retrieval and semantic English information retrieval. This is achieved by two sub-processes of the search engine: translation of the original or expanded query to the target language and shallow semantic linguistic inference, facilitating the retrieval of documents which contain phrases that imply the meaning of the (translated) query, even when no exact match of the keywords is found.

These enhancements are facilitated via a specification of the domain.
Setting Up a Basic Retrieval Application

- A basic retrieval system was created by
  - Collecting relevant documents
    - An archaeology expert searched the Web for relevant sites (in Hebrew and English)
    - All the documents linked from those pages were collected using a crawler
    - In total, we collected a non-comparable bilingual corpus for Archaeology containing several thousand documents in English and Hebrew
  - Implementing a search engine over the collected documents
    - All documents were indexed using the basic Jakarta Lucene indexing and search engine
    - Instead of inflected words, we indexed the lemma of each word
    - In order to match the indexed terms, query terms are also lemmatized
We use TEASE (Szpektor et al., 2004), a state-of-the-art unsupervised acquisition algorithm for lexical-syntactic entailment rules.

TEASE acquires entailment relations for a given input template from the Web:

- Retrieve sentences that match the input template.
- From these sentences, extract anchor-sets, which are identified as being characteristic for the input template based on statistical criteria.
- Retrieves from the Web sentences that contain the extracted anchor-sets.
- Parse the retrieved sentences and replace the anchors found in each sentence with their corresponding variables.
- From this retrieved corpus of parsed sentences, learn templates that are assumed to entail or be entailed by the input template.
TEASE learns entailment rules for a given input template

To acquire rules, an archaeology expert generated a list of domain-specific verbs and verb phrases

We then executed TEASE on each of the templates representing these verbs in order to learn from the Web rules in which the input templates participate

We learned approximately 3900 rules for 80 input templates
### Input Template vs. Learned Template

| Input Template          | Learned Template                                                                 |
|-------------------------|----------------------------------------------------------------------------------|
| X excavate Y            | X discover Y, X find Y, X uncover Y, X examine Y, X unearth Y, X explore Y        |
| X construct Y           | X build Y, X develop Y, X create Y, X establish Y                                 |
| X contribute to Y       | X cause Y, X linked to Y, X involve in Y                                          |
| date X to Y             | X built in Y, X began in Y, X go back to Y                                        |
| X cover Y               | X bury Y, X provide coverage for Y                                               |
| X invade Y              | X occupy Y, X attack Y, X raid Y, X move into Y                                   |
| X restore Y             | X protect Y, X preserve Y, X save Y, X conserve Y                                 |
Query expansion

- Queries are parsed with the Minipar dependency parser (Lin, 1998)
- They are then matched against the left hand side template of every rule in the learned knowledge-base
- Whenever a match is found, a new query is generated, in which the constant terms of the matched left hand side template are replaced with the constant terms of the right hand side template
- For example, given the query “excavations of Jerusalem by archaeologists” and a learned rule “excavation of Y by X → X dig in Y”, a new query is generated, containing the terms “archaeologists dig in Jerusalem”
Semantic expansion

1) Israel Antiquities Authority - Gallery of Sites and Finds
   id 99, score: 1.1237792521715164
   matched query: boat find | boat date | boat excavator

   Snippet:
   In 1985-86, the level of the Sea of Galilee dropped considerably owing to a drought, and the mud bottom of the lake off the former shoreline became exposed.

2) Ancient Egypt: Ships and Boats
   id 1600, score: 1.086829513311386
   matched query: boat uncover | boat find remains

   Snippet:
   Ancient Egyptian ships and boats: The archaeological evidence, the state involvement, ship construction, sailing the ships, constructions facilitating navigation, piracy
Cross-lingual IR

- Cross-lingual language technology is utilized in three different components of the system:
  - Hebrew documents are morphologically processed to provide better indexing
  - Query terms in English are translated to Hebrew and vice versa;
  - Hebrew snippets are translated to English
We use the HAMSAH morphological analyzer (Yona and Wintner, 2007) to process the entire domain-specific corpus. Resulting lexemes are used to index documents. This pre-processing brought to the foreground several omissions of the analyzer, mostly due to domain-specific terms missing in the lexicon. We selected the one thousand most frequent words with no morphological analysis and added their lexemes to the lexicon.
Query translation

- When users submit a query in one language they are provided with the option to request a translation of the query to the other language, thereby retrieving documents in the other language.
- In order to support cross-lingual query specification we capitalized on a medium-size bilingual dictionary.
- Since the coverage of the dictionary was rather limited, and many domain-specific items were missing, we chose the one thousand most frequent lexemes which had no translations and translated them manually.
- In order to translate query terms we use the Hebrew-English dictionary also as an English-Hebrew dictionary.
When Hebrew documents are retrieved, the (Hebrew) snippet which the system produces is augmented by an English translation.

We use an extended, improved version of a rudimentary Hebrew to English MT system (Lavie et al., 2004).

Extensions include an improved morphological analysis of the input, an extended bilingual dictionary and a revised set of transfer rules, as well as a more modern transfer engine and a much larger language model for generating the target (English) sentences.

Domain specific adaptation of the system is relatively easy, and does not require a domain specific parallel corpus.
Translation example

A JAR OF THE ANCIENT BRONZE PERIOD WITH OYSTERS FROM THE NILWS
Conclusions

- A system for cross-lingual and semantically-enhanced retrieval of information in the cultural heritage domain
- An instance of adapting existing state-of-the-art tools and resources to the domain
- Quantitative evaluation is hard
- A preliminary analysis of a sample of queries shows improvement
### Results

| Query                              | Without Expansion | With Expansion |
|------------------------------------|-------------------|---------------|
|                                    | Relevant in Top 10 | Relevant in Top 10 | Total Retrieved | Total Retrieved |
| discovering boats                  | 2                 | 2             | 5             | 86             |
| growing vineyards                  | 0                 | 0             | 6             | 8              |
| Persian invasions                  | 5                 | 5             | 8             | 22             |
| excavations of the Byzantine period | 10                | 37            | 10            | 100            |
| restoring mosaics                  | 0                 | 0             | 3             | 69             |

The number of relevant documents out of the top 10 and the total number of retrieved documents (up to 100) for some queries.
The system learned lexical syntactic rules that cannot be expressed by a mere lexical substitution:
- “date X to Y ↔ X go back to Y”

Rules are not necessarily restricted to learning from the target domain:
- the rule “X discover Y ↔ X find Y” was learned from contexts such as \{X=“astronomers”; Y=“new planets”\} and \{X=“zoologists”; Y=“new species”\}

Applying semantic query expansion, followed by English to Hebrew query translation, results in query expansion for Hebrew using techniques that were so far applicable only to resource-rich languages.
References

Bar-Haim, Roy, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In Second PASCAL Challenge Workshop for Recognizing Textual Entailment.

Dagan, Ido, Oren Glickman, and Bernardo. Magnini. 2006. The pascal recognising textual entailment challenge. In Lecture Notes in Computer Science, Volume 3944, volume 3944, pages 177–190.

Lavie, Alon, Shuly Wintner, Yaniv Eytani, Erik Peterson, and Katharina Probst. 2004. Rapid prototyping of a transfer-based Hebrew-to-English machine translation system. In Proceedings of TMI-2004: The 10th International Conference on Theoretical and Methodological Issues in Machine Translation, Baltimore, MD, October.

Lin, Dekang. 1998. Dependency-based evaluation of minipar. In Proceedings of the Workshop on Evaluation of Parsing Systems at LREC 1998.

Lytinen, S., N. Tomuro, and T. Repede. 2000. The use of wordnet sense tagging in faqfinder. In Proceedings of the AAAI00 Workshop on AI and Web Search.

Szpektor, Idan, Hristo Tanev, Ido Dagan, and Bonaventura Coppola. 2004. Scaling web-based acquisition of entailment relations. In Proceedings of EMNLP 2004.

Yona, Shlomo and Shuly Wintner. 2007. A finite-state morphological grammar of Hebrew. Natural Language Engineering. To appear.

Zukerman, Ingrid and Bhavani Raskutti. 2002. Lexical query paraphrasing for document retrieval. In Proceedings of ACL 2002.