Cross Lingual and Semantic Retrieval for Cultural Heritage Appreciation

Idan Szpektor, Ido Dagan
Dept. of Computer Science
Bar Ilan University
szpekti@cs.biu.ac.il

Alon Lavie
Language Technologies Inst.
Carnegie Mellon University
alavie+@cs.cmu.edu

Danny Shacham, Shuly Wintner
Dept. of Computer Science
University of Haifa
shuly@cs.haifa.ac.il

Abstract

We describe a system which enhances the experience of museum visits by providing users with language-technology-based information retrieval capabilities. The system consists of a cross-lingual search engine, augmented by state of the art semantic expansion technology, specifically designed for the domain of the museum (history and archaeology of Israel). We discuss the technology incorporated in the system, its adaptation to the specific domain and its contribution to cultural heritage appreciation.

1 Introduction

Museum visits are enriching experiences: they provide stimulation to the senses, and through them to the mind. But 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, either on location or elsewhere. One common means of exploration is Information Retrieval (IR) via a Search Engine. For example, a museum could implement a search engine over a collection of documents relating to the topics exhibited in the museum.

However, such document collections are usually much smaller than general collections, in particular the World Wide Web. Consequently, phenomena inherent to natural languages may severely hamper the performance of human language technology when applied to small collections. One such phenomenon is the semantic *variability* of natural languages, the ability to express a specific meaning in many different ways. For example, the expression *"Archaeologists found a new tomb"* can be expressed also by *"Archaeologists discovered a tomb"* or *"A sarcophagus was dug up by Egyptian Researchers"*. On top of monolingual variability, the same information can also be expressed in different languages. Ignoring natural language variability may result in lower recall of relevant documents for a given query, especially in small document collections.

This paper describes a system that attempts to cope with semantic variability through the use of state of the art human language technology. The system provides both semantic expansion and cross lingual IR (and presentation of information) in the domain of archaeology and history of Israel. It was specifically developed for the Hecht Museum in Haifa, Israel, which contains a small but unique collection of artifacts in this domain. The system provides different users with different capabilities, bridging over language divides; it addresses semantic variation in novel ways; and it thereby complements the visit to the museum with long-lasting instillation of information.

The main component of the system is 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; and cross-lingual expansion based on translation of Hebrew queries to English and vice versa. Retrieved documents are presented as links with associated snippets; the system also translates snippets from Hebrew to English.

The main contribution of this work is, of course, the system itself, which was recently demonstrated

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 65–72,
Prague, 28 June 2007.

successfully at the museum and which we believe could be useful to a variety of museum visitor types, from children to experts. For example, the system provides Hebrew speakers access to English documents pertaining to the domain of the museum, and vice versa, thereby expanding the availability of multilingual material to museum visitors. 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, thereby improving their accuracy and usability. 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.

2 Background and Motivation

Internet search is hampered by the complexity of natural languages. The two main characteristics of this complexity are *ambiguity* and *variability*: the former refers to the fact that a given text can be interpreted in more than one way; the latter indicates that the same meaning can be linguistically expressed in several ways. The two phenomena make simple search techniques too weak for unsophisticated users, as existing search engines perform only direct keyword matching, with very limited linguistic processing of the texts they retrieve.

Specifically, IR systems that do not address the variability in languages may suffer from lower recall, especially in restricted domains and small document locations. We next describe two prominent types of variability that we think should be addressed in IR systems.

2.1 Textual Entailment and Entailment Rules

In many NLP applications, such as Question Answering (QA), Information Extraction (IE) and Information Retrieval (IR), it is crucial to recognize that a specific target meaning can be inferred from different text variants. For example, a QA system needs to induce that *"Mendelssohn wrote incidental music"* can be inferred from *"Mendelssohn composed incidental music"* in order to answer the question *"Who wrote incidental music?"*. This type of reasoning has been identified as a core semantic inference task by the generic *textual entailment* framework (Dagan et al., 2006; Bar-Haim et al., 2006).

The typical way to address variability in IR is to use lexical query expansion (Lytinen et al., 2000; Zukerman and Raskutti, 2002). However, there are variability patterns that cannot be described using just constant phrase to phrase entailment. 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 compose Y $\rightarrow$ X write Y'. The left hand side is assumed to entail the right hand side in certain contexts, under the same variable instantiation. Paraphrases can be viewed as bidirectional entailment rules. Such rules capture basic inferences in the language, and are used as building blocks for more complex entailment inference. For example, given the above entailment rule, a QA system can identify the answer *"Mendelssohn"* in the above example. This need sparked intensive research on automatic acquisition of paraphrase and entailment rules.

Although knowledge-bases of entailment-rules and paraphrases learned by acquisition algorithms were used in other NLP applications, such as QA (Lin and Pantel, 2001; Ravichandran and Hovy, 2002) and IE (Sudo et al., 2003; Romano et al., 2006), to the best of our knowledge the output of such algorithms was never applied to IR before.

2.2 Cross Lingual Information Retrieval

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. For example, 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 relatively poor at searching for them. Other visitors may be unable to read Hebrew texts, but still benefit from non-textual information that are contained in Hebrew documents (e.g., pictures, maps, audio and video files, external links, etc.)

This problem is addressed by the paradigm of Cross-Lingual Information Retrieval (CLIR). This paradigm has become a very active research area in recent years, addressing the needs of multilingual and non-English speaking communities, such as the

European Union, East-Asian nations and Spanish speaking communities in the US (Hull and Grefenstette, 1996; Ballesteros and Croft, 1997; Carbonell et al., 1997). The common approach for CLIR is to translate a query in a source language to another target language and then issue the translated query to retrieve target language documents. As explained above, CLIR research has to address various generic problems caused by the variability and ambiguity of natural languages, as well as specific problems related to the particular languages being addressed.

3 Coping with Semantic Variability in IR

We describe a search engine that is capable of performing: (a) semantic English information retrieval; and (b) cross-lingual (Hebrew-English and English-Hebrew) information retrieval, allowing users to pose queries in either of the two languages and retrieve documents in both. This is achieved by two sub-processes of the search engine: first, the engine performs shallow semantic linguistic inference and supports the retrieval of documents which contain phrases that imply the meaning of the translated query, even when no exact match of the translated keywords is found. This is enabled by automatic acquisition of semantic variability patterns that are frequent in the language, which extend traditional lexical query expansion techniques. Second, the engine translates the original or expanded query to the target language, based on several linguistic processes and a machine readable bilingual dictionary. The result is a semantic expansion of a given query to a variety of alternative wordings in which an answer to this query may be expressed in the target language of the retrieved documents.

These enhancements are facilitated via a specification of the domain. As our system is specifically designed to work in the domain of the history and archaeology, we could focus our attention on resources and tools that are dedicated to this domain. Thus, for example, lexicons and dictionaries, whose preparation is always costly and time consuming, were developed with the specific domain in mind; and textual entailment and paraphrase patterns were extracted for the specific domain. While the resulting system is focused on visiting the Hecht Museum, the methodology which we used and discuss here can be adapted to other areas of cultural heritage, as well as to other narrow domains, in the same way.

3.1 Setting Up a Basic Retrieval Application

We created a basic retrieval system in two steps: first, we collected relevant documents; then, we implemented a search engine over the collected documents.

In order to construct a local corpus, an archaeology expert searched the Web for relevant sites and pages. We then downloaded all the documents linked from those pages using a crawler. The expert looked for documents in both English and Hebrew. In total, we collected a non-comparable bilingual corpus for Archaeology containing several thousand documents in English and Hebrew.

We implemented our enhanced retrieval modules on top of the basic Jakarta Lucene indexing and search engine[1]. All documents were indexed using Lucene, but instead of inflected words, we indexed the lemma of each word (see detailed description of our Hebrew lemmatization in Section 3.3). In order to match the indexed terms, query terms (either Hebrew or English) were also lemmatized before the index was searched, in a manner similar to lemmatizing the documents.

3.2 Query Expansion Using Entailment Rules

As described in Section 2.1, entailment rules had not been used as a knowledge resource for expanding IR queries, prior to our work. In this paper we use this resource instead of the typical lexical expansion in order to test its benefit. Most entailment rules capture relations between different predicates. We thus focus on documents retrieved for queries that contain a predicate over one or two entities, which we term here *Relational IR*. We would like to retrieve only documents that describe an occurrence of that predicate, but possibly in words different than the ones used in the query. In this section we describe in detail how we learn entailment rules and how we apply them in query expansion.

Automatically Learning Entailment Rules from the Web Many algorithms for automatically learning paraphrases and entailment rules have been explored in recent years (Lin and Pantel, 2001;

[1] http://jakarta.apache.org/lucene/docs/index.html

Ravichandran and Hovy, 2002; Shinyama et al., 2002; Barzilay and Lee, 2003; Sudo et al., 2003; Szpektor et al., 2004; Satoshi, 2005). In this paper 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. It first retrieves from the Web sentences that match the input template. From these sentences it extracts the variable instantiations, termed *anchor-sets*, which are identified as being characteristic for the input template based on statistical criteria.

Next, TEASE retrieves from the Web sentences that contain the extracted anchor-sets. The retrieved sentences are parsed and the anchors found in each sentence are replaced with their corresponding variables. Finally, from this retrieved corpus of parsed sentences, templates that are assumed to entail or be entailed by the input template are learned. The learned templates are ranked by the number of occurrences they were learned from.

Entailment Rules for Domain Specific Query Expansion Our goal is to use the knowledge-base of entailment rules learned by TEASE in order to perform query expansion. The two subtasks that arise are: (a) acquiring an appropriate knowledge-base of rules; and (b) expanding a query given such a knowledge-base.

TEASE learns entailment rules for a given input template. As our document collection is domain specific, a list of such relevant input templates can be prepared. In our case, we used an archaeology expert to generate a list of verbs and verb phrases that relate to archaeology, such as: *'excavate'*, *'invade'*, *'build'*, *'reconstruct'*, *'grow'* and *'be located in'*. 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. An example for such rules is presented in Table 1. We learned approximately 3900 rules for 80 input templates.

Since TEASE learns lexical-syntactic rules, we need a syntactic representation of the query. We parse each query using the Minipar dependency parser (Lin, 1998). We next try to match the left hand side template of every rule in the learned knowledge-base. Since TEASE does not identify the direction of the relation learned between two templates, we try both directional rules that are induced from a learned relation. 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 $X \rightarrow X$ dig in Y'*, a new query is generated, containing the terms *'archaeologists dig in Jerusalem'*. Finally, we retrieve all the documents that contain all the terms of at least one of the expanded queries (including the original query). The basic search engine provides a score for each document. We re-score each document as the sum of scores it obtained from the different queries that it matched. Figure 1 shows an example of our query expansion, where the first retrieved documents do not contain the words used to describe the predicate in the query, but other ways to describe it.

All the templates learned by TEASE contain two variables, and thus the rules that are learned can only be applied to queries that contain predicates over two terms. In order to broaden the coverage of the learned rules, we automatically generate also all the partial templates of a learned template. These are templates that contain just one of variables in the original template. We then generate rules between these partial templates that correspond to the original rules. With partial templates/rules, expansion for the query in Figure 1 becomes possible.

3.3 Cross-lingual IR

Until very recently, linguistic resources for Hebrew were few and far between (Wintner, 2004). The last few years, however, have seen a proliferation of resources and tools for this language. In this work we utilize a relatively large-scale lexicon of over 22,000 entries (Itai et al., 2006); a finite-state based morphological analyzer of Hebrew that is directly linked to the lexicon (Yona and Wintner, 2007); a medium-size bilingual dictionary of some 24,000 word pairs; and a rudimentary Hebrew to English machine translation system (Lavie et al., 2004). All these resources had to be adapted to the domain of the Hecht museum.

Cross-lingual language technology is utilized in

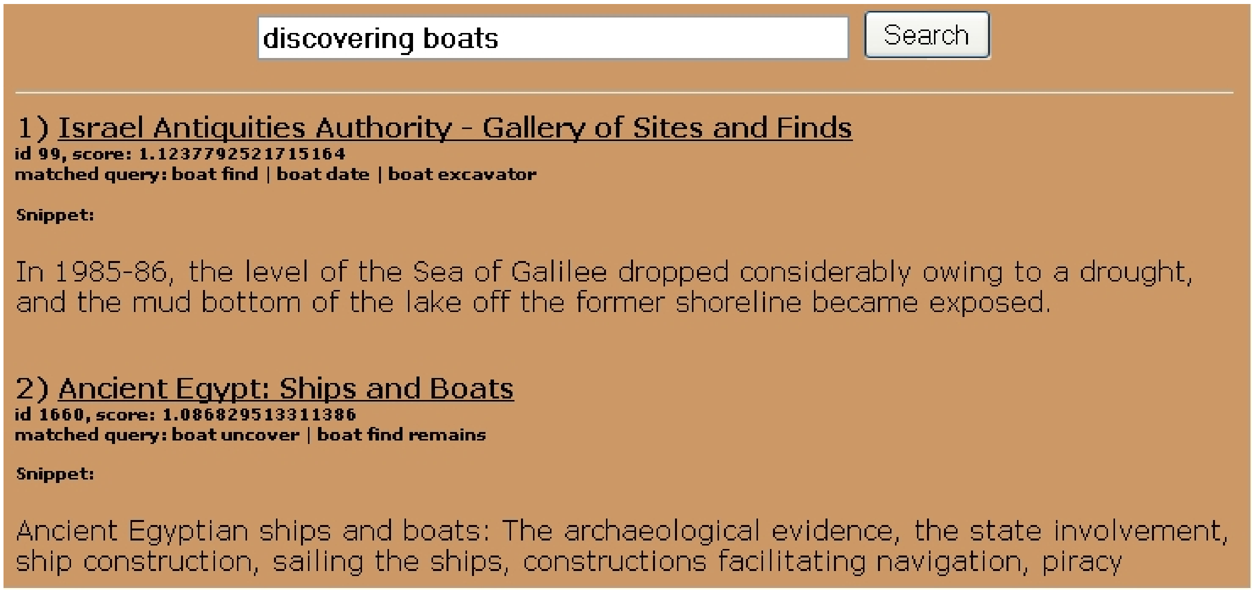

Figure 1: Semantic expansion example. Note that the expanded queries that were generated in the first two retrieved texts (listed under '**matched query**') do not contain the original query.

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; and Hebrew snippets are translated to English. We discuss each of these components in this section.

Linguistically-aware indexing The correct level of indexing for morphologically-rich language has been a matter of some debate in the information retrieval literature. When Arabic is concerned, Darwish and Oard (2002) conclude that "Character n-grams or lightly stemmed words were found to typically yield near-optimal retrieval effectiveness". Since Hebrew is even more morphologically (and orthographically) ambiguous than Arabic, and especially in light of the various prefix particles which can be attached to Hebrew words, we opted for full morphological analysis of Hebrew documents before they are indexed, followed by indexing on the lexeme.

We use the HAMSAH morphological analyzer (Yona and Wintner, 2007), which was recently rewritten in Java and is therefore more portable and efficient (Wintner, 2007). We processed the entire domain specific corpus described above and used the resulting lexemes to index documents. This preprocessing 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. While we do not have quantitative evaluation metrics, the coverage of the system improved in a very evident way.

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. The motivation behind this capability is that users who may be able to read documents in a language may find the specification of queries in that language too challenging; also, retrieving documents in a foreign language may be useful due to the non-textual information in the retrieved documents, especially in a museum environment.

In order to support cross-lingual query specification we capitalized on a medium-size bilingual dictionary that was already used for Hebrew to English machine translation. 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 transla-

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

Table 1: Examples for correct templates that were learned by TEASE for input templates.

tions and translated them manually, augmenting the lexicon with missing Hebrew lexemes where necessary and expanding the bilingual dictionary to cover this domain.

In order to translate query terms we use the Hebrew English dictionary also as an English-Hebrew dictionary. While this is known to be sub-optimal, our current results support such an adaptation in lieu of dedicated directional bilingual dictionaries.

Translating a query from one language to another may introduce ambiguity where none exists. For example, the query term *spinh* 'vessel' is unambiguous in Hebrew, but once translated into English will result in retrieving documents on both senses of the English word. Usually, this problem is overcome since users tend to specify multi-term queries, and the terms disambiguate each other. However, a more systematic solution can be offered since we have access to semantic expansion capabilities (in a single language). That is, expanding the query in the source language will result in more query terms which, when translated, are more likely to disambiguate the context. We leave such an extension for future work.

Snippet translation When Hebrew documents are retrieved, we augment the (Hebrew) snippet which the system produces by an English translation. We use an extended, improved version of a rudimentary Hebrew to English MT system developed by 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.

The MT system is transfer based: it performs linguistic pre-processing of the source language (in our case, morphological analysis) and post-processing of the target (generation of English word forms), and uses a small set of transfer rules to translate local structures from the source to the target and create translation hypotheses, which are stored in a lattice. A statistical language model is used to decode the lattice and select the best hypotheses.

The benefit of this architecture is that domain specific adaptation of the system is relatively easy, and does not require a domain specific parallel corpus (which we do not have). The system has access to our domain-specific lexicon and bilingual dictionary, and we even refined some transfer rules due to peculiarities of the domain. One advantage of the transfer-based approach is that it enables us to treat out-of-lexicon items in a unique way. We consider such items proper names, and transfer rules process them as such. As an example, Figure 2 depicts the translation of a Hebrew snippet meaning *A jar from the early bronze period with seashells from the Nile*. The word *nilws* 'Nile' is missing from the lexicon, but this does not prevent the system from producing a legible translation, using the transliterated form where an English equivalent is unavailable.

4 Conclusions

We described a system for cross-lingual and semantically-enhanced retrieval of information in the cultural heritage domain, obtained by adapting existing state-of-the-art tools and resources to the domain. The system enhances the experience of museum visits, using language technology as a vehicle for long-lasting instillation of information. Due to the novelty of this application and the dearth of available multilingual annotated resources in this domain, we are unable to provide a robust, quan-

Figure 2: Translation example

Query	**Without Expansion**		**With Expansion**	
	Relevant in Top 10	Total Retrieved	Relevant in Top 10	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

Table 2: Analysis of the number of relevant documents out of the top 10 and the total number of retrieved documents (up to 100) for a sample of queries.

titative evaluation of the approach. A preliminary analysis of a sample of queries is presented in Table 2. It illustrates the potential of expansion for document collections of narrow domain. In what follows we provide some qualitative impressions.

We observed that the system was able to learn many expansion rules that cannot be induced from manually constructed lexical resources, such as thesauri or WordNet (Fellbaum, 1998). This is especially true for rules that are specific for a narrow domain, e.g. 'X restore Y $\rightarrow$ X preserve Y'. Furthermore, the system learned lexical syntactic rules that cannot be expressed by a mere lexical substitution, but include also a syntactic transformation. For example, 'date X to Y $\leftrightarrow$ X go back to Y'.

In addition, since rules are acquired by searching the Web, they are not necessarily restricted to learning from the target domain, but can be learned from similar terminology in other domains. For example, the rule 'X *discover* Y $\leftrightarrow$ X *find* Y' was learned from contexts such as $\{X$='*astronomers*';Y='*new planets*'$\}$ and $\{X$='*zoologists*';Y='*new species*'$\}$.

The quality of the rules that were automatically acquired is mediocre. We found that although many rules were useful for expansion, they had to be manually filtered in order to retain only rules that achieved high precision.

Finally, we note that applying semantic query expansion (using entailment rules), 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, such as English.

Acknowledgements

This research was supported by the Israel Internet Association; by THE ISRAEL SCIENCE FOUNDATION (grant No. 137/06 and grant No. 1095/05); by the Caesarea Rothschild Institute for Interdisciplinary Application of Computer Science at the University of Haifa; by the ITC-irst/University of Haifa collaboration; and by the US National Science Foundation (grants IIS-0121631, IIS-0534217, and the Office of International Science and Engineering).

We wish to thank the Hebrew Knowledge Center at the Technion for providing resources for Hebrew. We are grateful to Oliviero Stock, Martin Golumbic, Alon Itai, Dalia Bojan, Erik Peterson, Nurit Melnik, Yaniv Eytani and Noam Ordan for their help and support.

References

Lisa Ballesteros and W. Bruce Croft. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 84–91.

Roy Bar-Haim, 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*.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL*.

Jaime G. Carbonell, Yiming Yang, Robert E. Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. 1997. Translingual information retrieval: A comparative evaluation. In *IJCAI (1)*, pages 708–715.

Ido Dagan, 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.

Kareem Darwish and Douglas W. Oard. 2002. Term selection for searching printed Arabic. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 261–268, New York, NY, USA. ACM Press.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press.

D. A. Hull and G. Grefenstette. 1996. Querying across languages. a dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th ACM SIGIR Conference*, pages 49–57.

Alon Itai, Shuly Wintner, and Shlomo Yona. 2006. A computational lexicon of contemporary Hebrew. In *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC-2006)*.

Alon Lavie, 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.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. In *Natural Language Engineering*, volume 7(4), pages 343–360.

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

S. Lytinen, 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*.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of ACL*.

Lorenza Romano, Milen Kouylekov, Idan Szpektor, Ido Dagan, and Alberto Lavelli. 2006. Investigating a generic paraphrase-based approach for relation extraction. In *Proceedings of EACL*.

Sekine Satoshi. 2005. Automatic paraphrase discovery based on context and keywords between ne pairs. In *Proceedings of IWP*.

Yusuke Shinyama, Satoshi Sekine, Sudo Kiyoshi, and Ralph Grishman. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of HLT*.

Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman. 2003. An improved extraction pattern representation model for automatic ie pattern acquisition. In *Proceedings of ACL*.

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

Shuly Wintner. 2004. Hebrew computational linguistics: Past and future. *Artificial Intelligence Review*, 21(2):113–138.

Shuly Wintner. 2007. Finite-state technology as a programming environment. In Alexander Gelbukh, editor, *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2007)*, volume 4394 of *Lecture Notes in Computer Science*, pages 97–106, Berlin and Heidelberg, February. Springer.

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

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