# Statistical Thesaurus Construction for a Morphologically Rich Language

**Chaya Liebeskind, Ido Dagan and Jonathan Schler**
Computer Science Department
Bar-Ilan University
Ramat-Gan, Israel
liebchaya@gmail.com, dagan@cs.biu.ac.il, schler@gmail.com

## Abstract

Corpus-based thesaurus construction for Morphologically Rich Languages (MRL) is a complex task, due to the morphological variability of MRL. In this paper we explore alternative term representations, complemented by clustering of morphological variants. We introduce a generic algorithmic scheme for thesaurus construction in MRL, and demonstrate the empirical benefit of our methodology for a Hebrew thesaurus.

## 1 Introduction

Corpus-based thesaurus construction has been an active research area (Grefenstette, 1994; Curran and Moens, 2002; Kilgarriff, 2003; Rychly and Kilgarriff, 2007). Typically, two statistical approaches for identifying semantic relationships between words were investigated: first-order, co-occurrence-based methods which assume that words that occur frequently together are topically related (Schutze and Pederson, 1997) and second-order, distributional similarity methods (Hindle, 1990; Lin, 1998; Gasperin et al, 2001; Weeds and Weir, 2003; Kotlerman et al., 2010), which suggest that words occurring within similar contexts are semantically similar (Harris, 1968).

While most prior work focused on English, we are interested in applying these methods to MRL. Such languages, Hebrew in our case, are characterized by highly productive morphology which may produce as many as thousands of word forms for a given root form.

Thesauri usually provide *related terms* for each entry term (denoted *target term*). Since both target and related terms correspond to word lemmas, statistics collection from the corpus would be most directly applied at the lemma level as well, using a morphological analyzer and tagger (Linden and Piitulainen, 2004; Peirsman et al., 2008; Rapp, 2009). However, due to the rich and challenging morphology of MRL, such tools often have limited performance. In our research, the accuracy of a state-of-the-art modern Hebrew tagger on a cross genre corpus was only about 60%.

Considering such limited performance of morphological processing, we propose a schematic methodology for generating a co-occurrence based thesaurus in MRL. In particular, we propose and investigate three options for term representation, namely surface form, lemma and multiple lemmas, supplemented with clustering of term variants. While the default lemma representation is dependent on tagger performance, the two other representations avoid choosing the right lemma for each word occurrence. Instead, the multiple-lemma representation assumes that the right analysis will accumulate enough statistical prominence throughout the corpus, while the surface representation solves morphological disambiguation "in retrospect", by clustering term variants at the end of the extraction process. As the methodology provides a generic scheme for exploring the alternative representation levels, each corpus and language-specific tool set might yield a different optimal configuration.

## 2 Methodology

Thesauri usually contain thousands of entries, termed here *target terms*. Each entry holds a list of *related terms*, covering various semantic relations. In this paper we assume that the list of target terms



*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 59–64,
Montréal, Canada, June 7-8, 2012. 

is given as input, and focus on the process of extracting a ranked list of candidate related terms (termed *candidate terms*) for each target term. The top ranked candidates may be further examined (manually) by a lexicographer, who will select the eventual related terms for the thesaurus entry.

Our methodology was applied for statistical measures of first order similarity (word co-occurrence). These statistics consider the number of times each candidate term co-occurs with the target term in the same document, relative to their total frequencies in the corpus. Common co-occurrence metrics are *Dice coefficient* (Smadja et al, 1996), *Pointwise Mutual Information* (PMI) (Church and Hanks, 1990) and *log-likelihood test* (Dunning, 1993).

## 2.1 Term Representation

Statistical extraction is affected by term representation in the corpus. Usually, related terms in a thesaurus are lemmas, which can be identified by morphological disambiguation tools. However, we present two other approaches for term representation (either a target term or a candidate related term), which are less dependent on morphological processing.

Typically, a morphological analyzer produces all possible analyses for a given token in the corpus. Then, a *Part Of Speech* (POS) tagger selects the most probable analysis and solves morphology disambiguation. However, considering the poor performance of the POS tagger on our corpus, we distinguish between these two analysis levels. Consequently, we examined three levels of term representation: (i) Surface form (*surface*) (ii) Best lemma, as indentified by a POS tagger (*best*), and (iii) All possible lemmas, produced by a morphological analyzer (*all*).

## 2.2 Algorithmic Scheme

We used the following algorithmic scheme for thesaurus construction. Our input is a target term in one of the possible term representations (*surface*, *best* or *all*). For each target term we retrieve all the documents in the corpus where the target term appears. Then, we define a set of candidate terms that consists of all the terms that appear in all these documents (this again for each of the three possible term representations). Next, a co-occurrence score between the target term and each of the candidates is calculated. Then, candidates are sorted, and the highest rated candidate terms are clustered into lemma-oriented clusters. Finally, we rank the clusters according to their members' co-occurrence scores and the highest rated clusters become related terms in the thesaurus.

Figure 1 presents the algorithm's pseudo code. The notion *rep*(term) is used to describe the possible term representations and may be either *surface*, *best* or *all*. In our experiments, when *rep*(target_term)=*best*, the correct lemma was manually assigned (assuming a lexicographer involvement with each thesaurus entry in our setting). While, when *rep*(word)=*best*, the most probable lemma is assigned by the tagger (since there are numerous candidates for each target term we cannot resort the manual involvement for each of them). The two choices for *rep*(term) are independent, resulting in nine possible configurations of the algorithm for representing both the target term and the candidate terms. Thus, these 9 configurations cover the space of possibilities for term representation. Exploring all of them in a systematic manner would reveal the best configuration in a particular setting.

```
Input: target term, corpus, a pair of values for
rep(target_term) and rep(word)
Output: clusters of related terms

target_term ← rep(target_term)
docs_list ← search(target_term)
FOR doc IN docs_list
   FOR word IN doc
      add rep(word) to candidates
   ENDFOR
ENDFOR
compute co-occurrence scores for all candidates
sort(candidates) by score
clusters ← cluster(top(candidates))
rank(clusters)
related terms ← top(clusters)
```

Figure 1: Methodology implementation algorithm

## 2.3 Clustering

The algorithm of Figure 1 suggests clustering the extracted candidates before considering them for the thesaurus. Clustering aims at grouping together related terms with the same lemma into clusters, using some measure of morphological equivalence. Accordingly, an equivalence measure between related terms needs to be defined, and a clustering

algorithm needs to be selected. Each obtained cluster is intended to correspond to the lemma of a single candidate term. Obviously, clustering is mostly needed for surface-level representation, in order to group all different inflections of the same lemma. Yet, we note that it was also found necessary for the lemma-level representations, because the tagger often identifies slightly different lemmas for the same term.

The equivalence measure is used for building a graph representation of the related terms. We represented each term by a vertex and added an edge between each pair of terms that were deemed equivalent. We investigated alternative equivalence measures for measuring the morphological distance between two vertices in our graph. We considered the string edit distance measure and suggested two morphological-based equivalence measures. The first measure, given two vertices' terms, extracts all possible lemmas for each term and searches for an overlap of at least one lemma. The second measure considers the most probable lemma of the vertices' terms and checks whether these lemmas are equal. The probability of a lemma was defined as the sum of probabilities for all morphological analyses containing the lemma, using a morpho-lexical context-independent probabilities approximation (Goldberg et al., 2008). The clustering was done by finding the connected components in our graph of terms using the JUNG[1] implementation (WeakComponentVertexClusterer algorithm with default parameters). The connected components are expected to correspond to different lemmas of terms. Hierarchical clustering methods (Jain et al., 1999) were examined as well (Single-link and Complete-link clustering), but they were inferior.

After applying the clustering algorithm, we re-ranked the clusters aiming to get the best clusters at the top of clusters list. We investigated two scoring approaches for cluster ranking; maximization and averaging. The maximization approach assigns the maximal score of the cluster members as the cluster score. While the averaging approach assigns the average of the cluster members' scores as the cluster score. The score obtained by either of the approaches may be scaled by the cluster length, to account for the accumulative impact of all class members (corresponding to morphological variants of the candidate term).

## 3 Case Study: Cross-genre Hebrew Thesaurus

Our research targets the construction of a cross genre thesaurus for the Responsa project[2]. The corpus includes questions posed to rabbis along with their detailed rabbinic answers, consisting of various genres and styles. It contains 76,710 articles and about 100 million word tokens, and was used for previous IR and NLP research (Choueka, 1972; Fraenkel, 1976; Choueka et al., 1987; Kernel et al, 2008).

Unfortunately, due to the different genres in the Responsa corpus, available tools for Hebrew processing perform poorly on this corpus. In a preliminary experiment, the POS tagger (Adler and Elhadad, 2006) accuracy on the Responsa Corpus was less than 60%, while the accuracy of the same tagger on modern Hebrew corpora is ~90% (Bar-Haim et al., 2007).

For this project, we utilized the MILA Hebrew Morphological Analyzer[3] (Itai and Wintner, 2008; Yona and Wintner, 2008) and the (Adler and Elhadad 2006) POS tagger for lemma representation. The latter had two important characteristics: The first is flexibility- This tagger allows adapting the estimates of the prior (context-independent) probability of each morphological analysis in an unsupervised manner, from an unlabeled corpus of the target domain (Goldberg et al., 2008). The second advantage is its mechanism for analyzing unknown tokens (Adler et al., 2008). Since about 50% of the words in our corpora are unknown (with respect to MILA's lexicon), such mechanism is essential.

For statistics extraction, we used Lucene[4]. We took the top 1000 documents retrieved for the target term and extracted candidate terms from them. Dice coefficient was used as our co-occurrence measure, most probable lemma was considered for clustering equivalence, and clusters were ranked based on maximization, where the maximal score was multiplied by cluster size.

[1] http://jung.sourceforge.net/

[2] Corpus kindly provided - http://www.biu.ac.il/jh/Responsa/

[3] http://mila.cs.technion.ac.il/mila/eng/tools_analysis.html

[4] http://lucene.apache.org/

## 4 Evaluation

### 4.1 Dataset and Evaluation Measures

The results reported in this paper were obtained from a sample of 108 randomly selected terms from a list of 5000 terms, extracted from two publicly available term lists: the University of Haifa's entry list[5] and Hebrew Wikipedia entries[6].

In our experiments, we compared the performance of the alternative 9 configurations by four commonly used IR measures: precision (P), relative recall (R), F1, and Average Precision (AP). The scores were macro-averaged. We assumed that our automatically-generated candidate terms will be manually filtered, thus, recall becomes more important than precision. Since we do not have any pre-defined thesaurus, we evaluated the relative-recall. Our relative-recall considered the number of suitable related terms from the output of all methods as the full set of related terms. As our system yielded a ranked sequence of related terms clusters, we also considered their ranking order. Therefore, we adopted the recall-oriented AP for ranking (Voorhees and Harman, 1999).

### 4.2 Annotation Scheme

The output of the statistical extraction is a ranked list of clusters of candidate related terms. Since manual annotation is expensive and time consuming, we annotated for the gold standard the top 15 clusters constructed from the top 50 candidate terms, for each target term. Then, an annotator judged each of the clusters' terms. A cluster was considered as relevant if at least one of its terms was judged relevant[7].

### 4.3 Results

Table 1 compares the performance of all nine term representation configurations. Due to data sparseness, the lemma-based representations of the target term outperform its surface representation. However, the best results were obtained from candidate representation at the surface level, which was complemented by grouping term variants to lemmas in the clustering phase.

| Target \ Candidate | | surface | best | All |
|---|---|---|---|---|
| **Surface** | R | 36.59 | 29.37 | 26.68 |
| | P | 24.29 | 21.09 | 18.71 |
| | F1 | 29.20 | 24.55 | 21.99 |
| | AP | 20.87 | 15.83 | 14.13 |
| **Best lemma** | R | 46.70 | 39.88 | 36.97 |
| | P | **25.03** | 23.08 | 20.94 |
| | F1 | **32.59** | 29.24 | 26.74 |
| | AP | 26.84 | 20.86 | 19.32 |
| **All lemmas** | R | **47.13** | 42.52 | 42.13 |
| | P | 23.72 | 22.47 | 21.23 |
| | F1 | 31.56 | 29.40 | 28.24 |
| | AP | **27.86** | 22.99 | 21.14 |

Table 1: Performances of the nine configuratrions

Furthermore, we note that the target representation by all possible lemmas (all) yielded the best R and AP scores, which we consider as most important for the thesaurus construction setting. The improvement over the common default best lemma representation, for both target and candidate, is notable (7 points) and is statistically significant according to the two-sided Wilcoxon signed-rank test (Wilcoxon, 1945) at the 0.01 level for AP and 0.05 for R.

## 5 Conclusions and Future Work

We presented a methodological scheme for exploring alternative term representations in statistical thesaurus construction for MRL, complemented by lemma-oriented clustering at the end of the process. The scheme was investigated for a Hebrew cross-genre corpus, but can be generically applied in other settings to find the optimal configuration in each case.

We plan to adopt our methodology to second order distributional similarity methods as well. In this case there is an additional dimension, namely feature representation, whose representation level should be explored as well. In addition, we plan to extend our methods to deal with Multi Word Expressions (MWE).

## Acknowledgments

This work was partially supported by the PASCAL-2 Network of Excellence of the European Community FP7-ICT-2007-1-216886.

[5] http://lib.haifa.ac.il/systems/ihp.html
[6] http://he.wikipedia.org
[7] This was justified by empirical results that found only a few clusters with some terms judged positive and others negative

## References


Adler Meni and Michael Elhadad. 2006. An Unsupervised Morpheme-Based HMM for Hebrew Morphological Disambiguation, in *Proceedings of COLING-ACL*, Sydney, Australia.

Adler Meni, Yoav Goldberg, David Gabay and Michael Elhadad. 2008. Unsupervised Lexicon-Based Resolution of Unknown Words for Full Morphological Analysis, in *Proceedings of ACL*.

Bar-Haim Roy, Khalil Sima'an, and Yoad Winter. 2007. Part-of-speech tagging of Modern Hebrew text. *Natural Language Engineering*, 14(02):223.251.

Choueka, Yaacov. 1972. Fast searching and retrieval techniques for large dictionaries and concordances. *Hebrew Computational Linguistics*, 6:12–32, July.

Choueka, Y., A.S. Fraenkel, S.T. Klein and E. Segal. 1987. Improved techniques for processing queries in full-text systems. *Proceedings of the 10th annual international ACM SIGIR conference on Research and development in information retrieval*.

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics* 16(1): 22–29.

Curran, James R. and Marc Moens. 2002. Improvements in automatic thesaurus extraction. In *Proceedings of the ACL-SIGLEX Workshop on Unsupervised Lexical Acquisition*, pages 59–67, Philadelphia, PA.

Dunning, T.E. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19, 61–74 (1993).

Fraenkel, Aviezri S. 1976. All about the Responsa retrieval project – what you always wanted to know but were afraid to ask. *Jurimetrics Journal*, 16(3):149–156, Spring.

Gasperin, C., Gamallo, P., Agustini, A., Lopes, G., and de Lima, V. 2001. Using syntactic contexts for measuring word similarity. In *the Workshop on Semantic Knowledge Acquisition and Categorisation (ESSLI 2001)*, Helsinki, Finland.

Goldberg Yoav, Meni Adler and Michael Elhadad, 2008. EM Can Find Pretty Good HMM POS-Taggers (When Given a Good Start), in *Proceedings of ACL*.

Grefenstette, G. 1994. Explorations in Automatic Thesaurus Construction. *Kluwer Academic Publishers*, Boston, USA

Harris, Zelig S. 1968. *Mathematical Structures of Language*. John Wiley, New York.

Hindle, D. 1990. Noun classification from predicate argument structures. In *Proceedings of ACL*.

Itai Alon and Shuly Wintner. 2008. Language Resources for Hebrew. *Language Resources and Evaluation* 42(1):75-98, March 2008.

Jain, A. K., M. N. Murty, P. J. Flynn. 1999. Data Clustering: A Review. *ACM Computing Surveys* 31(3):264-323.

Kerner Yaakov HaCohen, Ariel Kass, Ariel Peretz. 2008. Combined One Sense Disambiguation of Abbreviations. In *Proceedings of ACL (Short Papers)*, pp. 61-64.

Kilgarriff, Adam. 2003. Thesauruses for natural language processing. In *Proceedings of the Joint Conference on Natural Language Processing and Knowledge Engineering*, pages 5–13, Beijing, China.

Kotlerman Lili, Dagan Ido, Szpektor Idan, and Zhitomirsky-Geffet Maayan. 2010. Directional Distributional Similarity for Lexical Inference. *Natural Language Engineering*, 16(4):359–389.

Linden Krister, and Jussi Olavi Piitulainen. 2004. Discovering Synonyms and Other Related Words. In Proceedings of COLING 2004 : CompuTerm 2004: 3rd International Workshop on Computational Terminology, Pages 63-70, Geneva, Switzerland

Lin, D. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*.

Peirsman Yves, Kris Heylen, and Dirk Speelman. 2008. Putting things in order. first and second order contexts models for the calculation of semantic similarity. In Actes des 9i`emes Journ´ees internationales d'Analyse statistique des Donn´ees Textuelles (JADT 2008), pages 907–916.

Rapp, R. 2009. The Automatic Generation of Thesauri of Related Words for English, French, German, and Russian, *International Journal of Speech Technology*, 11, 3-4, 147-156.

Rychly, P. and Kilgarriff, A. 2007. An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In *Proceedings of ACL-07*, demo session. Prague, Czech Republic.

Schutze Hinrich and Jan 0. Pederson. 1997. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management*, 33(3):307-318.

Smadja, F., McKeown, K.R., Hatzivassiloglou, V. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics* 22, 1–38

Voorhees E.M. and D. Harman. 1999. Overview of the seventh text retrieval conference . In *Proceedings of the Seventh Text Retrieval 73 Conference*, 1999. NIST Special Publication. 58.

Weeds, J., and Weir, D. 2003. A general framework for distributional similarity. In *Proceedings of EMNLP*, Sapporo, Japan.

Wilcoxon F. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83.

Yona Shlomo and Shuly Wintner. 2008. A Finite-State Morphological Grammar of Hebrew. *Natural Language Engineering* 14(2):173-190, April 2008. *Language Resources and Evaluation* 42(1):75-98, March 2008.