Acronym-Expansion Recognition and Ranking on the Web

Alpa Jain
Columbia University
alpa@cs.columbia.edu

Silviu Cucerzan
Microsoft Research
silviu@microsoft.com

Saliha Azzam
Microsoft
salihaa@microsoft.com

Abstract

The paper presents a study on large-scale automatic extraction of acronyms and associated expansions from Web data and from the user interactions with this data through Web search engines. We investigate three information sources for extracting *and* ranking *acronym-expansion pairs, as provided by a large-scale search engine: the crawled web documents, the search engine logs, and the search results. We evaluate and compare the acronym-expansion pairs generated from these sources on three dimensions: (1) the precision and recall of each source; (2) the overlap and inclusion among the acronym-expansion sets; and (3) the rank-order correlation of the ordered expansion sets. Our results show that all three data sources play an important role in building a comprehensive up-to-date collection of acronym-expansion pairs.*

1 Introduction

Acronyms are commonly defined as abbreviations or short descriptors of phrases, formed from the initial letters of the important terms in a phrase. We refer to the originating phrases as *expansions*. For example, "NAACL" is often used as an abbreviation for "the North American chapter of the Association for Computational Linguistics", while "AMVET" is used as an abbreviation for the expansion "American Veterans". In some work, the term acronym is used only to describe pronounceable abbreviations, but such distinction is beyond the scope of this work. While relatively new, the use of acronyms has become a widespread linguistic phenomenon, to the point that almost any combination of four or fewer letters is employed as an acronym nowadays. Moreover, the same acronym may have different expansions depending on the context in which it occurs.

Information retrieval tasks can highly benefit from equating acronyms to their expansions: one such benefit is expanding the pool of documents considered at query processing time for user queries that contain acronyms or expansions associated with an acronym. For instance, the phrase "Human Immunodeficiency Virus" is more commonly referred using the acronym "HIV", e.g.,"HIV vaccine". For a search query such as "human immunodeficiency virus vaccine," the documents that contain "HIV vaccine" may not be considered as a match if they do not contain the query terms also. In such situations, substituting "HIV" for "human immunodeficiency virus" allows expanding the set of documents to be returned as results, and possibly improving result quality. Other information-retrieval-related tasks such as query expansion and query suggestion can also benefit from a relation consisting of acronym-expansion pairs.

As we show further, when extracting acronyms from the Web, we obtain not only a large number of acronyms, but also a very large number of expansions per acronym. For example, we are able to extract no less than 133 expansions for the acronym "ABC" from web pages used in our experiments. In such situations, the information retrieval scenario of interchanging acronyms and their expansions becomes impractical without a structural change of the indexing strategy of the search engine. Thus, an important task, not addressed previously, is to rank the expansions.

Our goal is to perform a detailed study of the extraction of acronyms and corresponding expansions from the Web, based on the three representative sources of information provided by a large-scale web search engine: the crawled documents, the queries sent by users to the search engine, and the top results retrieved by the search engine. Towards this goal, we designed three methods to automatically extract acronyms-expansion pairs from each of the sources. Post extraction, each acronym-expansion pair is scored using an appropriate ranking algorithm for each source.

The first method mines text documents on the Web to generate the acronym-expansion relation, and scores each pair based on factors such as co-occurrence, popularity, and reliability. The second method focuses on the use of acronyms by the users of the search engine, viz, frequently searched acronyms. Specifically, we exploit the query refinement information that can be extracted from the query logs of a search engine to identify pairs of successive queries that contain an acronym and a possible expansion. The intuition behind this approach is that given the widespread use of acronyms on the Web and the convenience of typing only a few characters, users tend to query the abbreviated form of the term they are interested in when such an abbreviation exists. In some of these cases, users then requery using the expansion, either because the search engine was not able to retrieve documents about the desired expansion or because the documents retrieved, although on topic, were not satisfactory for the user need. In other cases, users are satisfied with the results and do not requery the expansion and thus, the acronym-expansion pair

Rank	Web documents	Query logs	Search results
1	american automobile association	american automobile association	american automobile association
2	automobile association of america	automobile association of america	amateur astronomers association
3	archives of american art	american auto association	american ambulance association
4	appraisers association of america	american arbitration association	american accordionists association
5	area agencies on aging	american automotive association	american accounting association
6	american avalanche association	arkansas activities association	automotive for all your automotive
7	american accounting association	abdominal aortic aneurysm	archives of american art
8	american arbitration association	agricultural adjustment act	
9	american association of anatomists	american airlines arena	
10	australian airports association	american academy of audiology	
1	american bonsai society	american breeders service	american bureau of shipping
2	american bamboo society	albino black sheep	australian bureau of statistics
3	american brachytherapy society	american bureau of shipping	the american bonanza society
4	atlas business solutions	allen b schwartz	the american budgerigar society
5	american board of surgery	amniotic band syndrome	acrylonitrile butadiene styreneplastic
6	automated bond system	american breeder service	amity business school
7	associated builder solutions	american bible society	ashland bus system
8	alternative behavioral services	american board of surgery	
9	american bladesmith society	american building supply	
10	asset backed securities	alternative behavioral services	

Table 1. Top 10 expansions for acronyms "AAA" and "ABS" generated using proposed sources.

is not captured by the query logs. This motivated a third approach for extracting acronym-expansion pairs, which uses the acronyms obtained from the logs as a starting point and processes the top search results retrieved by the search engine for those acronyms to extract the expansions available among the results and are possibly missing from the logs.

We present a comparison of the acronym-expansion relations that were generated and ranked using the three sources mentioned above. Interestingly, while there is a substantial overlap among these relations, each method extracts a large number of pairs not produced by the other methods (e.g., Table 1 illustrates this using ten top-ranked expansions generated by each source for "AAA" and "ABS"). We discuss the extent of correlation and disagreement among the ranking produced by each method. To the best of our knowledge, no other work has been carried out to study such diverse sources to extract and rank acronym-expansion pairs.

The paper is structured as follows: we first describe our approach to extract and rank acronym-expansion pairs using three different information sources. Then, we present our experimental evaluation of individual relations as well as comparisons of the proposed methods.

2 Extracting Acronym-Expansion Pairs from a Web Crawl

The main advantage of extracting acronyms from web documents, instead of relying on specialized databases or acronym sites, is the ability to deal with the growing amount of textual information in web documents, email, and newsgroup data, where new acronyms are created every day, before dictionaries, encyclopedia or specialized acronym collections can be updated. To process the text data available from a web crawl, we designed a two stage extraction approach based on existing techniques for acronym-expansion extraction. First, we identify candidate sentences using a finite-state lexical scanner and then we examine these candidate sentences for any acronym-expansion pairs.

To identify a candidate sentence, we follow the approach in [1, 9, 11] and apply simple character matching heuristics such as checking for uppercase letters and parentheses. In addition to these heuristics, we use linguistic features by identifying simple phrase structures, using a lightweight part-of-speech tagger, to allow for prepositional phrases as in "ACMC (Assistant Commandant of the Marine Corps)." Once a candidate sentence is identified, we search for acronym-expansion pairs using further heuristics suggested in [5, 7] which mainly focus on the length and the order of the letters in an acronym and the words in its expansion. Intuitively, these heuristics restrict the number of stopwords (determiners, prepositions, and conjunctions) and *significant words* (non-stopwords) in an expansion. These constraints can be tuned based on the precision and recall requirements: if precision is crucial, the number of content words in the expansion can be restricted to not exceed the total number of letters in the acronym. This constraint will select the expansion "Certified Public Accounting" in "Candidate for Certified Public Accounting (CPA)", correctly excluding the word "candidate", but may reject good expansions such as "National Historic Landmarks" in "National Historic Landmarks Program (NHL)".

Finally, we carry out *expansion normalization* to identify semantically identical expansions with non-identical string representations (e.g., "Frequently Asked Question" and "Frequently-Asked Questions"). We detect such variations using a rule-based approach that compares expansions while ignoring punctuations and the grammatical number (singular or plural) of the terms in the expansion phrases. To select a canonical representation, we rely on ranks assigned to these expansions, as described later in Section 5.

We processed about 4 million arbitrary documents, after filtering out HTML tags, from the web crawl data for Live

Search[1]. The extraction and ranking process was carried out on a single machine, over a period of few days, and generated 176,190 pairs.

3 Extracting Acronym-Expansion Pairs from Search Query Logs

While the acronym-expansion relation obtained from the web crawl provides insightful information about the usage of acronyms on the Web, it is not necessarily indicative of the acronym knowledge and usage in the general population of web users. To capture this component, we resorted to statistics extracted from logs of queries issued by users of a large scale, popular web search engine (Live Search). The expansions provided by Live Search users, in conjunction with the information supplied by the top search results, provide a snapshot of the users' perception and usage of acronyms.

Parsing search engine query logs to generate acronym-expansion pairs presents a number of challenges. Since all popular search engines, including the one employed, are case-insensitive, users seldom submit the queries cased properly and thus, query logs may not provide sufficient information about the proper capitalization of the words in the queries. In particular, the Live Search query logs contained all queries in lower case form. Furthermore, queries typically do not contain both an acronym and its expansion and therefore, the syntactic patterns used in existing work used to process web documents would fail to identify any acronym-expansion pairs.

To overcome these problems, we developed an algorithm to construct acronym-expansion pairs from web search sessions, by analyzing pair of queries that were submitted in succession by the search engine users. Specifically, we only processed pairs of queries in which:

- first query contains at least two characters and no spaces;
- second query contains at least five characters;
- first query is not a substring of the second query.

For these pairs, we perform a dynamic programming match in which we allow consecutive letters in the first query to match word prefixes of length up to three in the second query with an option of skipping stop words. This strategy allowed us to match correctly pairs such as "cool" and "cooperation in ontology and linguistics", which a simple greedy match strategy would fail to recognize.

We analyzed a Live Search query log containing almost 100 million query pairs and generated more than 50,000 acronym-expansion pairs, which indicates a relatively high popularity of acronym-related queries. Since query logs are plagued by misspellings even more than the web documents, we employed the spell checker of the search engine to remove the misspelled expansions, to ultimately obtain 43,413 acronym-expansion pairs.

[1]http://www.live.com

4 Extracting Acronym-Expansion Pairs from Search Results

While the search-log-based strategy studies the acronym-expansion pairs popular among the search-engine users, it disregards the fact that some expansions may have been present in the top search results and the users did not have to refine their queries. To compensate for this effect, we employed a method to extract acronyms from the top search result returned by a popular web search engine.

We started with the list of acronyms generated from the query log mining process, and issued each acronym as a search query to Live Search. We retrieved the top 50 search results, from which we extracted the document titles. We then employed the matching strategy based on search-session query pairs with one modification, allowing matching the acronym to a substring of the title (rather than the whole title); for example, the acronym "AAA" can match the first three words of the retrieved title "Amateur Astronomers Association of New York City". While the simple presence of an acronym in a document or a candidate expansion in the title does not provide guarantees about the relationship between the document and the acronym, such a relationship is very likely to exist for documents retrieved in the top 50 results by a web search engine. Note also that this strategy may lead to cropping out parts of the official name of the entity targeted by the document (e.g. "of New York City"), which can be detrimental to our goal, as shown by the precision numbers in our experiments, but can also be regarded as a positive generalization effect.

5 Ranking the Acronym Expansions

So far, we described our extraction process for each of the three information sources. We observed that each method generates a large number of acronyms with multiple associated expansions, e.g., we retrieved 55 expansions for "AAA" using the web crawl data (see Table 1). To discriminate among the expansions associated with an acronym, we assign a score to each expansion and then rank the expansions based on their scores. We now discuss the ranking of acronym expansions for each source.

Web Crawl: Given an acronym A and expansion E pair generated from the web crawl data (Section 2), we compute the *Score(A, E)* based on the following factors:

- Co-occurrence between A and E
- Popularity of the pair (A, E)
- Reliability of sources for the pair (A, E)

To estimate these factors, we rely on documents that contain both the acronym A and the expansion E, as retrieved by sending queries of the form "*expansion (acronym)*" to a web search engine.[2] We refer to these queries as *co-occurrence*

[2]We can also use "acronym (expansion)" but preliminary experiments showed that such queries do not modify the relative ranking.

queries. We compute the score of a pair after parsing the search results obtained for the co-occurrence queries.

Co-occurrence: We measure the association strength between an acronym and its expansion using the pointwise mutual information (PMI), computed as $log_2 \frac{P(A \text{ and } E)}{P(A) \cdot P(E)}$, where $P(A \text{ and } E)$ denotes the probability that A and E co-occur, and $P(A)$ and $P(E)$ denote the occurrence probabilities of A and E, respectively. These probabilities are estimated using the PMI-IR algorithm in [10], which employs the expected number of web hits as computed by a Web search engine for the queries "E" and "E(A)": ***PMI(A, E)*** $= \frac{hits(\text{“}E(A)\text{”})}{hits(E)}$. ($P(A)$ is common across the scores of all expansions.)

Popularity: Oftentimes, acronyms that are uncommon on the Web are created for local usage, e.g., within a community, organization, or a website. If these acronyms and their expansions are often mentioned together in such a domain, the co-occurrence value may be relatively high for such pairs. For this reason, we also measure the *popularity* of an acronym-expansion pair as the number of unique domains among the URLs returned in the results for the co-occurrence query, and use it as a correction factor.

Reliability: Finally, we want to boost the scores for expansions hosted on *reliable* web pages. Using *PageRank* as an indicator of reliability, we compute the reliability score of an acronym-expansion pair as the average PageRank of top k pages in the results for the co-occurrence query.[3]

After normalizing the number of unique domains and average PageRank, we compute the score of an acronym-expansion pair as: $Score(A, E) = PMI(A, E) \cdot D(A, E) \cdot S(A, E)$, where ***D(A, E)*** and ***S(A,E)*** is the normalized number of unique domains and average PageRank respectively, with 1 being the maximum score.

Query Logs: To rank expansions of an acronym generated from search query logs (Section 3), we use the frequency, i.e., the number of times an expansion follows an acronym in user search sessions.

Search Results: We investigated two methods to rank expansions extracted using the search results (Section 4): one in which the expansions are ranked using the highest ranked document from which they were extracted and another in which the ranks of all such generating documents are aggregated. Since these two methods returned similar results (in terms of the Spearman correlation of the ranked lists), we used the former as being the simpler of the two.

6 Findings

We first evaluated each of the methods by computing the *precision* and *recall* of the extracted relations. Additionally, we measured the *overlap* and *inclusion* of the relations and finally, we compared the three ranked expansion lists.

[3]We chose k=10 pages to keep the ranking time low.

6.1 Precision and Recall

A major challenge in evaluating precision of large size collections of acronym-expansion pairs such as those generated by our methods is that no comprehensive *gold standard* set exists. To compute the precision, we drew a random sample of 2,498 pairs and manually verified the correctness of the expansions provided by each method. Specifically, we issued the co-occurrence queries (see Section 5) for each acronym-expansion pair to Live Search and manually examined the results, including the document title, text, etc., for evidence that the pair is correct. We calculated the precision for a set of acronym-expansion pairs (E_s) as: $\frac{|\textit{Correct acronym-expansion pairs}|}{|E_s|}$. The web crawl-based method has the highest precision, of 96.7%, followed by the log-based method with 90% and the search-result-based method with 81.2%. When we combine all methods, by constructing a collective set of acronym-expansion pairs, the precision is 91.5%.

Method	Acronym recall	Expansion recall	Overall pairs
Crawl	0.8	0.45	176,190
Query logs	0.57	0.27	43,413
Search results	0.53	0.19	43,413
Combined	0.82	0.49	-

Table 2. Number of acronym-expansion pairs and recall of the three sources as calculated using Wikipedia.

To compute the recall, we generated a list of acronym-expansion pairs from Wikipedia, and used these pairs as the *reference set*. For each method, we measured the *acronym recall* and the *expansion recall*. Specifically, for a set of acronyms (A) extracted by a method and the set of acronyms in the reference set (R_a), we compute the acronym recall as $\frac{|A \cap R_a|}{|R_a|}$. To measure the expansion recall, we consider the acronyms common to both the reference set and an extracted relation ($A \cap R_a$). If E denotes the set of expansions extracted for acronyms in $A \cap R_a$ by one of our methods and R_e denotes the corresponding expansions in the reference set, then the expansion recall is computed as $\frac{|E \cap R_e|}{|R_e|}$. Table 2 lists the acronym and expansion recall for each method and the combination of the three.

The precision and recall numbers we obtained are consistent with those reported in previous work [9, 11, 5]. However, using the novel query-session-based sources show a drop in recall and precision (as expected) since they only focus on the frequently searched acronym-expansion pairs. However, these methods are orders of magnitude faster than the crawl-based methods and thus, our experiments quantify the quality-time tradeoff between the sources. Overall, our experiments show that suitability of each approach depends on the system requirements: if precision if critical, using the web-crawl-based method is preferable, whereas if recall is critical, it is beneficial to combine all three methods.

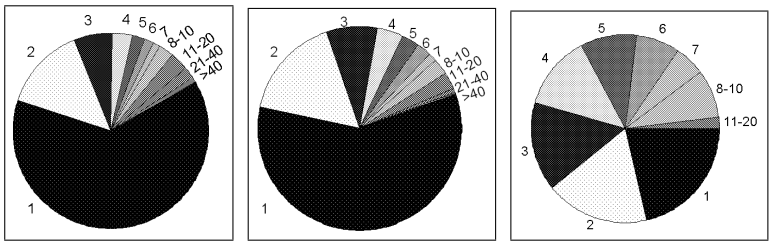

Figure 1. Number of expansions for acronyms extracted from: (a) web documents; (b) query logs; (c) search results.

6.2 Comparing The Proposed Methods

We compared the acronym-expansion relation generated by each method along multiple dimensions. Specifically, we studied the overall distribution of the number of expansions associated with an acronym in each relation as well as the pair-wise *overlap* and *inclusion* for the relations.

Figure 1 illustrates the percentage of extracted acronyms with a certain number of expansions, for each method. For instance, among the acronyms extracted from the web crawl data, 64% were associated with 1 expansion, 14% were associated with 2 expansions, and 1% exceeded 40 expansions. The distributions for the acronyms obtained from the query logs and the web crawl are quite similar. The large fraction associated with one expansion in both cases suggests that a substantial number of acronyms are used with only one meaning on the web and by the search engine users. In contrast, the search results seem to provide a much balanced distribution of possible expansions.

Reference	Test	Acronym recall	Expansion recall
Crawl	Logs	0.37	0.14
Crawl	Search	0.34	0.12
Search	Crawl	0.73	0.20
Search	Logs	1.00	0.15
Logs	Crawl	0.67	0.28
Logs	Search	0.84	0.20

Table 3. Inclusion across the acronym-expansion pairs generated by each method.

Overlap and Inclusion: To measure the degree of overlap between the three methods, we examine each pair of methods. Specifically, given a pair of acronym-expansion relations, we determine the number of expansions, for each acronym, that are common to both relations. Additionally, we calculate the *overlap ratio* by dividing the number of overlapping expansions by the minimum number of expansions for that acronym (which is the maximum achievable overlap). Figure 2 shows the pair-wise overlap as well as overlap ratios for all three methods. It is noteworthy that when the actual overlap is small, the maximum achievable overlap also tends to be small, which suggests that there are many acronyms with a small number of popular expansions that can be extracted from any of the three resources.

We further studied the pair-wise *inclusion* of acronym-expansion relations to identify whether any single relation completely subsumed the other relations. We measured the pair-wise inclusion, by holding one of the relations as the reference set and measuring the acronym and expansion recall as defined earlier. The inclusion values for all 6 pairs are shown in Table 3. Interestingly, while the acronym recall is high, the expansion recall indicates that the three methods are able to extract different expansions from the three web sources employed. In conjunction with the high precision numbers obtained, this suggests that a comprehensive list of expansions, which covers both the web usage and the user needs, cannot be obtained from mining (based on strict matching patterns) only one information source.

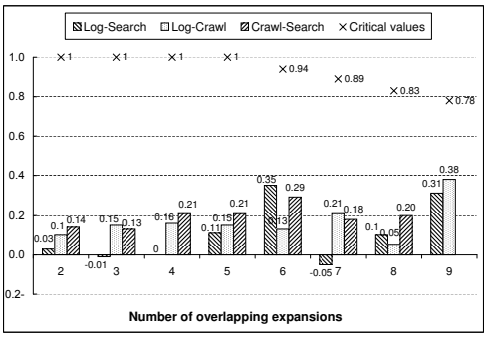

Figure 3. Average Spearman correlation coefficient among the three ranked lists of expansions.

Rank Correlation: Figure 3 synthesizes the results from the Spearman [3] rank correlation tests carried out on the ranked lists generated by each of the methods. The correlation coefficients, averaged over the overlapping expansion lists of various sizes, were overwhelmingly positive, but below the critical values and do not allow us to conclude that the rankings computed by these methods are correlated.

7 Related Work

Several approaches to automatically extract acronym-expansion pairs from text documents have been successfully investigated. [9] proposed AFP (Acronym Finding Program), which uses the longest common subsequence to identify acronyms of 3 to 10 uppercase letters formed from initials of expansion terms, ignoring stopwords. [11] proposed TLA (Three Letter Acronyms), which segments the text into *chunks* and checks adjacent chunks for candidate

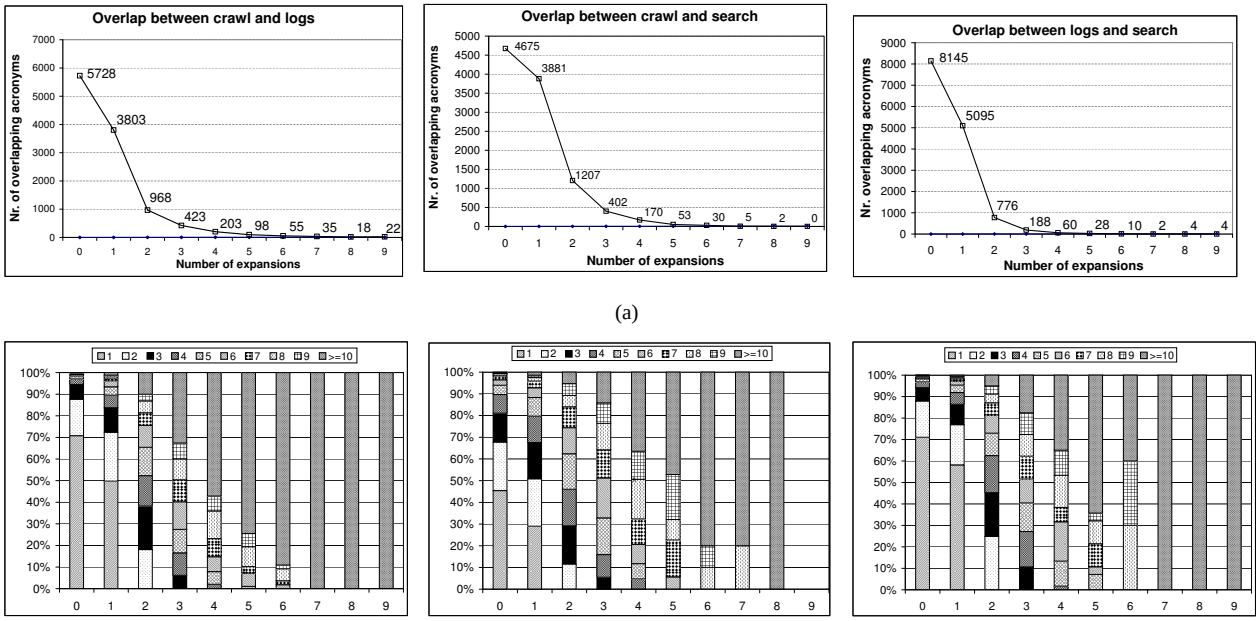

Figure 2. Overlap (a) and overlap ratios (b) between the three acronym-expansion pairs.

acronym-expansion pairs while allowing matches of up to three of the first letters in an expansion term. [6] used pattern-based rules along with text markers (e.g., '(..)' and '[..]') and cue terms (e.g.,"short", "stand") as indicators of acronym-expansion pairs. [13] segments the text based on lexical, morphological and phonological clues. Each segment can then contribute to the letters in an acronym. [5] uses a machine learning approach based on weak constraints instead of parenthetical and upper case expressions to increase the set of candidate pairs, and then employs supervised learning. [1] presented an approach based on logistic regression after restricting the candidate pairs to be of the form "expansion (acronym)". [2] extracted acronyms from Swedish text using machine learning algorithms on training data that is automatically generated by a rule-based algorithm. Other recent work [7, 8] has focused on extracting acronyms from biomedical texts, task which presents additional challenges. Also related to the general extraction task is the problem of disambiguating acronyms [7, 4, 12].

In general, existing work is targeted more towards designing methods to automatically identify acronym-extraction pairs in text documents. These approaches resembles our extraction method based on web crawl. However, in our work, we explore two additional sources available (directly or indirectly) from the Web, and propose automated methods to extract acronym-expansion relations from these sources. We further look into the problem of ranking expansions for an acronym, which is essential when dealing with large collections of acronym-expansion pairs.

8 Conclusion

We presented three methods for extracting ranked acronym-expansion relations from three different large data sources available to a search engine. These sources capture the main acronym usage on the Web, accounting for both the Web content creators and the Web search-engine users. Our experiments indicate that each method/source presents its strengths and that any comprehensive Web-based acronym extraction effort should employ all the sources investigated. While the acronym-expansion extraction from web documents exhibits the best precision, the methods based on mining query logs and search engine result sets capture to a better degree the current Web user needs and they can be employed to update any large collection of acronyms with up-to-the-minute information.

References

[1] J. Chang, H. Schutze, and R. Altman. Creating an online dictionary of abbreviations from medline. In *JAMIA*, 2002.

[2] D. Dannells. Automatic acronym recognition. In *EACL*, 2003.

[3] R. Hogg and A. Craig. *Introduction to Mathematical Statistics.* Macmillan, 1995.

[4] G. Kikui and E. Sumita. Using the Web to disambiguate acronyms. In *HLT-NAACL:Short papers*, 2006.

[5] Nadeau and P. Turney. A supervised learning approach to acronym identification. In *18th Canadian Conference on AI*, 2005.

[6] Y. Park and R. Byrd. Hybrid text mining for finding abbreviations and their definitions. In *EMNLP*, 2001.

[7] J. Pustejovsky et al. Extraction and disambiguation of acronym-meaning pairs in Medline. In *unpublished manuscript*, 2001.

[8] A. Schwartz et al. A simple algorithm for identifying abbreviation definitions in biomedical text. In *PSB*, 2003.

[9] K. Taghva and J. Gilbreth. Recognizing acronyms and their definitions. In *International Journal of Document Analysis and Recognition*, 1999.

[10] P. Turney. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *ECML*, 2001.

[11] S. Yeates. Automatic extraction of acronyms from text. In *NZ CSRSC*, 1999.

[12] M. Zahariev. Automatic sense disambiguation for acronyms. In *SIGIR*, 2004.

[13] M. Zahariev. A linguistic approach to extracting acronym expansions from text. In *KAIS*, 2004.