Organizing and Searching the World Wide Web of Facts - Step One: the One-Million Fact Extraction Challenge

Marius Paşca
Google Inc.
Mountain View, CA
mars@google.com

Dekang Lin
Google Inc.
Mountain View, CA
lindek@google.com

Jeffrey Bigham*
Univ. of Washington
Seattle, WA
jbigham@cs.washington.edu

Andrei Lifchits*
Univ. of British Columbia
Vancouver, BC
alifchit@cs.ubc.ca

Alpa Jain*
Columbia Univ.
New York, NY
alpa@cs.columbia.edu

Abstract

Due to the inherent difficulty of processing noisy text, the potential of the Web as a decentralized repository of human knowledge remains largely untapped during Web search. The access to billions of binary relations among named entities would enable new search paradigms and alternative methods for presenting the search results. A first concrete step towards building large searchable repositories of factual knowledge is to derive such knowledge automatically at large scale from textual documents. Generalized contextual extraction patterns allow for fast iterative progression towards extracting one million facts of a given type (e.g., Person-BornIn-Year) from 100 million Web documents of arbitrary quality. The extraction starts from as few as 10 seed facts, requires no additional input knowledge or annotated text, and emphasizes scale and coverage by avoiding the use of syntactic parsers, named entity recognizers, gazetteers, and similar text processing tools and resources.

Introduction

Motivation

A daily routine for hundreds of millions of Internet users, Web search provides simplified, yet practical keyword-based access to documents containing knowledge of arbitrary complexity levels. The potential of the Web as a repository of human knowledge is largely untapped during search, partly due to the inherent difficulty of representing and extracting knowledge from noisy natural-language text. Whereas full query and document understanding are distant, if not unfeasible goals, Web search can and should benefit from at least a small fraction of the implicit knowledge that is relatively easier to identify and extract from arbitrary Web documents.

A particularly useful type of knowledge for Web search consists in binary relations associated to named entities. The facts often occur in small text fragments "hidden" within much longer documents, e.g. the facts that *"the capital of Australia is Canberra"*, or *"Mozart was born in 1756"*, or *"Apple Computer is headquartered in Cupertino"*. A search engine with access to hundreds of millions of such Web-derived facts can answer directly fact-seeking queries, including fully-fledged questions and database-like queries (e.g., *"companies headquartered in Mountain View"*), rather than providing pointers to the most relevant documents that may contain the answers. Moreover, for queries referring to named entities, which constitute a large portion of the most popular Web queries, the facts provide alternative views of the search results, e.g., by presenting the *birth year* and *best-selling album* for singers, or *headquarters*, *name of CEO* and *stock symbol* for companies, etc.

Large-Scale Fact Extraction from the Web

Proposed in (Riloff & Jones 1999), the idea of unsupervised bootstrapping for information extraction was expanded and applied to the construction of semantic lexicons (Thelen & Riloff 2002), named entity recognition (Collins & Singer 1999), extraction of binary relations (Brin 1998; Agichtein & Gravano 2000), and acquisition of structured data for tasks such as Question Answering (Lita & Carbonell 2004; Cucerzan & Agichtein 2005; Fleischman, Hovy, & Echihabi 2003). In the same spirit, the approach introduced in this paper starts from a small set of seed items (in this case, facts), and iteratively grows it by finding contextual patterns that extract the seeds from the text collection, then identifying a larger set of candidate seeds that are extracted by the patterns, and adding a few of the best candidate seeds to the previous seed set. Our emphasis is on large-scale extraction, as a consequence of aggressive goals with respect to the amount of text to be mined, in the order of hundreds of millions of textual Web documents, and especially number of facts to be extracted, with a target of one million facts of a given (specific rather than general) type extracted with a precision of 80% or higher. Since exhaustive lists of hand-written extraction rules are domain specific and impractical to prepare, and large lists of seed facts are difficult to compile, the extraction must progress towards the final set of extracted facts starting from a seed set in the order of 10 facts. This corresponds to a growth rate of 100,000:1 between the size of the extracted set and the size of the initial set of seed facts. To our knowledge, the growth rate and the size of the extracted set of facts targeted here are several orders of magnitude higher than in any of the previous studies on fact extraction based on either hand-written extraction rules (Cafarella *et al.* 2005), or bootstrapping for relation and information extraction (Agichtein & Gravano 2000; Lita & Carbonell 2004).

*Contributions made during internships at Google Inc.

To overcome the limitations of previous work and achieve the coverage required by large-scale fact extraction, we introduce a method for the generalization of contextual extraction patterns. Traditionally, basic extraction patterns are slowly accumulated over hundreds or thousands of iterations - an approach that is inefficient, if not unfeasible, on Web-scale document collections. With pattern generalization, the iterative acquisition of facts progresses quickly by producing (much) higher-coverage extraction patterns in early iterations. By necessity, the resulting acquisition method aims for generality and lightweight processing. In comparison, besides the implicit reliance on clean text collections such as news corpora (Thelen & Riloff 2002; Agichtein & Gravano 2000; Hasegawa, Sekine, & Grishman 2004) rather than noisy Web documents, previous work often depends on text processing that is either relatively expensive, in the case of shallow syntactic parsing (Riloff & Jones 1999; Thelen & Riloff 2002), or restrictive to a small number of categories to which the facts could apply, in the case of named entity recognition (Agichtein & Gravano 2000).

Generalized Extraction Patterns

Basic Contextual Extraction Patterns

A seed fact is represented by the pair of phrases that are in a "hidden" relation, e.g. (*Vincenzo Bellini*, *1801*) for Person-BornIn-Year facts, (*Athens*, *Greece*) for City-CapitalOf-Country, and (*Portuguese*, *Brazil*) for Language-SpokenIn-Country. Following (Agichtein & Gravano 2000), each occurrence of the two sides of the fact within the same sentence produces a basic contextual extraction pattern, i.e., a triple (Prefix, Infix, Postfix). The prefix and postfix are contiguous sequences of a fixed number of terms, situated to the immediate left of the first matched phrase (e.g., to the left of *Athens*), and to the immediate right of the second matched phrase (e.g., to the right of *Greece*) respectively. The infix contains all terms between the two matched phrases (e.g., between *Athens* and *Greece*). For example, the occurrence of the seed fact (*Athens*, *Greece*) in the sentence *"The proposed trip will take students to Athens, the capital of Greece and home to 40% of the Greek population"* produces a pattern with the prefix [.. *take students to*], the infix [*, the capital of*] and the postfix [*and home to* ..].

For efficiency, the matching of seed facts onto sentences is implemented with a modified trie. Initially, both phrases of the seed facts are loaded into the trie. Each sentence is then matched onto the trie, resulting in a new contextual pattern only if both parts of the same seed fact could be successfully matched. The same process is applied without modifications later, during the acquisition of candidate facts, by matching triples (extraction patterns) onto sentences to extract pairs (new candidate facts). The matching over individual sentences is parallelized, following a programming model for processing large data sets (Dean & Ghemawat 2004).

Generalization via Distributionally Similar Words

The generalized patterns are produced from the basic extraction patterns. For that purpose, the terms in the prefix, infix and postfix of each basic pattern are replaced with their corresponding classes of distributionally similar words, if any. The classes are computed on the fly over the entire set of basic patterns, on top of a large set of pairwise similarities among words. The set of distributionally similar words (Lin 1998) is extracted in advance from around 50 million news articles indexed by the Google search engine over three years. All digits in both patterns and sentences are replaced with a marker (by convention, 0), such that any two numerical values with the same number of digits will overlap during matching.

Prefix	Infix	Postfix
CL3 00th :	's Birthday (	) . EndOfSent
StartOfSent	CL4 CL8 CL22 CL26 born	in CL17 ,
Memorial CL47 in	(b. CL3 0 ,	, d. CL3
among CL6 ...	CL4 born on 00 CL3	in CL10 ,
CL8 child :	CL4 born 00 CL3	in Lewisburg ,
CL4 written by	who CL4 born CL3 00 ,	, in Oak

CL3 = {March, October, April, Mar, Aug., February, Jul, Nov., ...}
CL4 = {is, was, has, does, could}
CL6 = {You, Lawmakers, Everyone, Nobody, Participants, ...}
CL8 = {a, the, an, each, such, another, this, three, four, its, most, ...}
CL10 = {Pennsylvania, Denver, Oxford, Marquette, Hartford, ...}
CL17 = {Tipperary, Rennes, Piacenza, Osasuna, Dublin, Crewe, ...}
CL22 = {Brazilian, Chinese, Japanese, Italian, Pakistani, Latin, ...}
CL26 = {entrepreneur, illustrator, artist, writer, sculptor, chef, ...}
.....
CL47 = {Tribute, Homage}
.....

Figure 1: Examples of generalized patterns acquired during the extraction of Person-BornIn-Year facts. A digit is represented by a 0.

In the process of replacing pattern words with word classes, some of the basic patterns become duplicates of one another. Therefore the set of generalized patterns is smaller than the set of basic patterns. However, the generalized patterns translate into significantly higher coverage than that of the basic patterns from which they were created. Consider the generalized patterns shown in Figure 1. The word classes CL4, CL8, CL22, CL26 and CL17 contain 5, 24, 87, 83 and 322 words respectively. After exhaustive enumeration of the elements in its word classes, the second generalization pattern alone from Figure 1 would be equivalent to an imaginary set of $5\times24\times87\times83\times322$=279,019,400 basic patterns. Even after discounting the imaginary basic patterns whose infixes are in fact unlikely or bogus sequences of words (e.g., *"does three Brazilian sculptor"*), this is a concrete illustration of why pattern generalization is important in large-scale extraction.

Validation and Ranking Criteria

It is possible to further increase the potential recall by a few orders of magnitude, after re-examining the purpose of the prefix and postfix in each generalized pattern. Strictly speaking, only the end of the prefix and the start of the postfix are useful, as they define the outer boundaries of a candidate fact. Comparatively, the actual word sequences in the pre-

fix and postfix introduce strong but unnecessary restrictions on the possible sentences that can be matched. A higher-recall alternative is to discard the prefix and postfix of all patterns. In their absence, the outer boundaries of the candidate facts in a sentence are computed in two steps. First, they are approximated through loose matching of the part-of-speech tags of sentence fragments to the left (e.g., [NNP NNP NNP] for *Robert S. McNamara*)) and to the right of the infix, on one hand, and the left (e.g., [NNP NNP] for *Stephen Foster*) and right sides of the seed facts, on the other hand. The resulting approximations are then validated, by verifying that the extremities of each candidate fact are distributionally similar to the corresponding words from one or more seed facts. For example, *Robert* is similar to *Stephen*, and *McNamara* is similar to *Foster*. Therefore, relative to a seed fact (*Stephen Foster*, *1826*), a new candidate fact (*Robert S. McNamara*, *1916*) is valid; comparatively, candidate facts such as (*Web Page*, *1989*) are discarded as invalid.

During validation, candidate facts are assigned similarity scores that aggregate individual word-to-word similarity scores of the component words relative to the seed facts. The similarity scores are one of a linear combination of features that induce a ranking over the candidate facts. Two other domain-independent features contribute to the final ranking, namely: a) a PMI-inspired (Turney 2001) score computed statistically over the entire set of patterns; the score promotes facts extracted by patterns containing words that are most indicative of the relation within the facts; and b) a completeness score computed statistically over the entire set of candidate facts, which demotes candidate facts if any of their two sides are likely to be incomplete (e.g., *Mary Lou* vs. *Mary Lou Retton*, or *John F.* vs. *John F. Kennedy*).

Experimental Setting

Text Collection: The source text collection consists of three chunks W_1, W_2, W_3 of approximately 100 million documents each. The documents are part of a larger snapshot of the Web taken in 2003 by the Google search engine. All documents are in English. The textual portion of the documents is cleaned of Html, tokenized, split into sentences and part-of-speech tagged using the TnT tagger (Brants 2000).

Target Facts: The evaluation involves facts of type Person-BornIn-Year. The reasons behind the choice of this particular type are threefold. First, many more Person-BornIn-Year facts are probably available on the Web (as opposed to, e.g., City-CapitalOf-Country facts), to allow for a good stress test for large-scale extraction. Second, either side of the facts (Person and Year) may be involved in many other types of facts, such that the extraction would easily diverge unless it performs correctly. Third, the phrases from one side (Person) have an utility in their own right, for applications related to lexicon construction or person name detection in Web documents.

Seed Set: The Person-BornIn-Year type is specified through an initial, randomly-selected set of 10 seed facts given as pairs: (*Irving Berlin*, *1888*), (*Hoagy Carmichael*, *1899*), (*Bob Dylan*, *1941*), (*Stephen Foster*, *1826*), (*John Lennon*, *1940*), (*Paul McCartney*, *1942*), (*Johann Sebastian Bach*, *1685*), (*Bela Bartok*, *1881*), (*Ludwig van Beethoven*, *1770*) and (*Vincenzo Bellini*, *1801*). Similarly to source documents, the facts are also part-of-speech tagged. No seed patterns are provided as input.

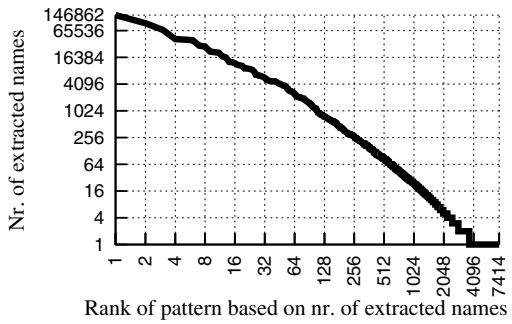

Figure 2: Number of unique names in the facts extracted by various extraction patterns from chunk W_1

Evaluation Results

Quantitative Results

An acquisition iteration consists in matching the seed facts onto sentences without crossing sentence boundaries, generating generalized acquisition patterns, and matching the patterns onto sentences to extract candidate facts. At the end of an iteration, approximately one third of the validated candidate facts are added to the current seed set. The acquisition expands the initial seed set of 10 facts to 100,000 facts (after iteration 1) and then to one million facts (after iteration 2) using chunk W_1. The 100,000 facts retained after iteration 1 generate a total of 89,186 generalized patterns, which correspond to 32,942 generalized patterns after discarding the prefix and postfix.

For the one million facts from W_1 placed in the seed set after iteration 2, Figure 2 illustrates the number of unique person names within the sets of facts extracted by various generalized infix-only patterns. The infixes of the patterns at rank 1, 2, 3 and 4 are the sequences [CL4 born in], [CL4 born on 00 CL3], [, b. CL3 00 ,] and [CL4 born 00 CL3] respectively (see Figure 1 for examples of words in classes CL3 and CL4).

The long tail of Figure 2 is also a characteristic of the number of unique person names within the sets of facts extracted from various Web hosts, as shown in Figure 3. One of the data points in the graph from Figure 3 corresponds to the Wikipedia encyclopedia (Remy 2002). It is the 342^{nd} most productive Web host, with birth years for only 339 people extracted from it. The relatively small number is encouraging, since chunk W_1 contains only a portion of the Wikipedia articles available in English as far back as 2003. Whereas there were approximately 750,000 articles in Wikipedia as of September 2005, just 570,000 were available only four months earlier, and clearly much fewer articles were crawled in 2003. The exact same patterns are likely to extract a larger number of facts, just by applying them to a Web snapshot containing more recent, and thus much larger, versions of Wikipedia and similar resources.

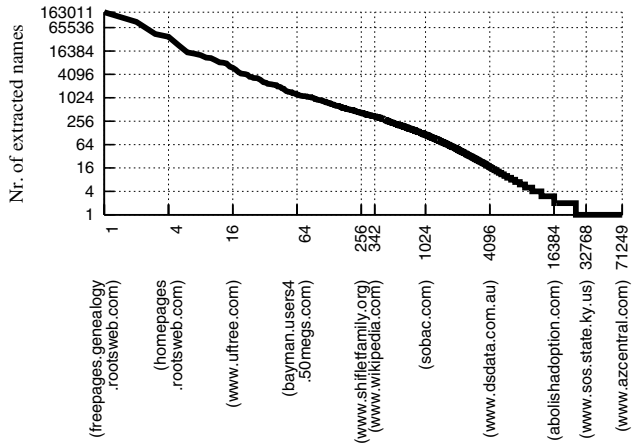

Figure 3: Number of unique names in the facts extracted from various hosts within chunk W_1

Accuracy of the Extracted Facts

Sample Set: For the purpose of evaluating precision, we select a sample of facts from the entire list of one million facts extracted from chunk W_1, ranked in decreasing order of their computed scores. The sample is generated automatically from the top of the list to the bottom, by retaining a fact and skipping the following consecutive N facts, where N is incremented at each step. The resulting list, which preserves the relative order of the facts, contains 1414 facts occurring at ranks $\{1, 2, 4, 7, 11, 16, .., 998992\}$. To better quantify precision at different ranks, the sample is then divided into 10 subsets of approximately 141 facts each. Some facts are discarded from the sample, if a Web search engine does not return any documents when the name (as a phrase) and the year are submitted together in a conjunctive query. In those cases, the facts were acquired from the 2003 snapshot of the Web, but queries are submitted to a search engine with access to current Web documents, hence the difference when some of the 2003 documents are no longer available or indexable. The first three columns of Table 1 illustrate the composition of the 10 subsets comprising the final sample for evaluating precision.

Evaluation Procedure: A fact is manually marked as correct only if there is clear evidence supporting it in some Web page. For obvious reasons, the search for such a Web page is done by submitting queries to a Web search engine, and freely perusing the returned documents. There are no a-priori constraints on the number of submitted queries or the documents being inspected. As expected, the evaluation is time-consuming; a fast human assessor needs more than two hours to classify 100 facts into correct and incorrect ones.

Several conditions must be verified manually before marking a fact as correct. The phrase on the left side of the fact must be that of a person, rather than a different type of entity such as (*Somerset Co.*, *1930*), (*Cover Letter*, *2000*) etc. Moreover, the name must be complete rather than partial or imprecise like (*Emma Jeane*, *1921*), (*Mrs. Huffman*, *1935*). On the other hand, the name should not contain extraneous words either; counterexamples include (*Enough Biography Desmond Llewelyn*, *1914*) and (*David Morrish Canadian*, *1953*). Finally and most importantly, the Web page should indicate clearly that the right side of the fact is the birth year of the person, rather than a different piece of information (e.g., *John Lennon*, *1980*). If no Web page is found that satisfies all these conditions, either because no such page exists or because the manual evaluation did not reach it, then the fact is marked as incorrect.

Table 1: Manual evaluation of precision over a sample of facts extracted from the first Web chunk (W_1)

Composition of Sample Set			Correct?		Precision
#	Rank Range	Facts	Yes	No	(%)
1	[1, 9871]	137	135	2	98.54
2	[10012, 39622]	125	118	7	94.40
3	[39904, 89677]	129	121	8	93.79
4	[90101, 159331]	130	118	12	90.76
5	[159896, 249572]	135	126	9	93.33
6	[250279, 359129]	133	125	8	93.98
7	[359977, 488567]	135	128	17	87.40
8	[489556, 639016]	130	92	38	69.38
9	[640147, 808357]	121	104	17	85.95
10	[809629, 998992]	124	93	31	75.00

Results: The precision values are shown in the fourth through sixth columns of Table 1. The average precision based on the sample set is 93.17% over approximately the top half of the list of one million facts (rank range [1, 488567]), and 88.38% over the entire list.

Coverage of the Extracted Facts

Gold Standard: The availability of external resources listing the birth dates for various people enables the pursuit of a fully automated, rather than manual evaluation of coverage. The gold standard set of facts is a random selection of 6617 pairs of people and their birth years from Wikipedia, including (*Abraham Louis Breguet*, *1747*), (*Clyde Tombaugh*, *1906*) and (*Quintino Sella*, *1827*).

Evaluation Procedure: The actual evaluation is automatic. It takes as input the facts in the gold standard, and a ranked list of extracted facts whose coverage is evaluated. For each fact from the gold standard, the evaluation consists of the following steps: a) identify the extracted facts with the same left side (corresponding to the person name), based on full case-insensitive string matching; b) collect the right side of the identified facts (i.e., the extracted birth years), if any, in the same relative order in which they occur in the ranked list of extracted facts; and c) compute the reciprocal rank score (Voorhees & Tice 2000) of the year from the gold-standard fact against the list of years identified in the previous step. Thus, if the list of years extracted for a given person name contains the correct year at rank 1, 2, 3, 4 or 5, the gold standard fact receives a score of 1, 0.5, 0.33, 0.25 or 0.2 respectively. The score is 0 if either none of the extracted facts matches the person name, or none of the years identified for the person name match the year from the gold standard. The overall score is the mean reciprocal rank

Table 2: Automatic evaluation of recall via MRR scores, over a random set of 6617 person names and their birth years

Evaluation Set Source and Size				Nr. Names with Some Extracted Year(s)						MRR Score	
W_1	W_2	W_3	Nr. Facts	@1	@2	@3	@4	@5	Bad	∩Gold	*All*Gold
√	-	-	1×10^6	2257	98	26	4	4	524	0.795	0.349
-	√	-	1×10^6	2095	80	23	4	7	528	0.784	0.324
-	-	√	1×10^6	2015	73	18	7	3	529	0.779	0.318
√	√	-	2×10^6	3049	163	34	11	13	574	0.819	0.475
√	√	√	3×10^6	3468	194	45	24	14	544	0.838	0.543

(MRR) score over multiple facts from the gold standard.

Results: Table 2 is a detailed view on the computed coverage values. The table consists of several vertical sections. The first section illustrates the source chunk(s) (one or more of W_1, W_2 or W_3) from which the evaluation set was extracted, and the number of facts in that set. The second section refers to the person names from the gold standard for which some extracted year exists in the evaluation set. That number is split into person names for which the correct year occurs at rank 1 through 5, as well as the person names for which all the extracted years are "Bad" since they are different from the year from the gold standard. For example, the first row of the table corresponds to the one million facts extracted and retained from chunk W_1. The facts provide the correct year at rank 1 for 2257 of the 6617 names from the gold standard, at rank 2 for 98, and at rank 5 for 4 of them, whereas some year(s) are extracted but they are all incorrect for 524 of the names.

The third section of Table 2 captures the average MRR scores computed only over the set of person names from the gold standard with some extracted year(s) (∩Gold), and then over the entire set of names from the gold standard (*All*Gold). The MRR score over ∩Gold measures, given that some year(s) were extracted for a person name, whether they include the year specified in the gold standard for that person name, and if so, how high that year is ranked relative to the other extracted years. Thus, if only one of the person names from the gold standard hypothetically occurred among the extracted facts, and the second year extracted for that name were correct, the MRR score over ∩Gold would be 0.5. On the other hand, the MRR score over *All*Gold is closer to the traditional definition of recall, although still more demanding. Indeed, an ideal recall of 1.0 requires all facts from the gold standard to be present anywhere in the evaluation set. In order for the MRR score over *All*Gold to reach the ideal value of 1.0, all facts from the gold standard must be present in the evaluation set, and the year from the gold standard must be the highest ranked year for that name within the evaluation set.

The size of the ∩Gold set of person names varies moderately, when the facts are extracted from chunk W_1 vs. W_2 vs. W_3. The same applies to MRR scores over ∩Gold and over *All*Gold. Comparatively, the size of the ∩Gold set increases significantly, as the evaluation sets merge facts extracted from more than one chunk. This translates into higher values for the MRR computed over *All*Gold. Note that the MRR score over the growing ∩Gold set remains stable, and even increases moderately with larger evaluation sets. The highest MRR score over the ∩Gold set is 0.838. Comparatively, the maximum MRR score over *All*Gold is 0.543.

For several reasons, the results in Table 2 are conservative assessments of the actual coverage. First, as noted earlier, the MRR scores are stricter, and lower, than corresponding recall values. Second, the presence of incorrect birth years in the gold standard, which cannot be extracted from the Web, is unlikely but possible given that anyone may volunteer to edit Wikipedia articles. Third, the use of full string matching, during the comparison of gold standard facts with extracted facts, generates artificially low scores (usually 0) for an unspecified number [1] of gold standard facts. Based on a quick post-evaluation scan, the most common cause of incorrect low scores seems to be a different degree of precision in specifying the person name, as shown by *William Shockley* vs. *William Bradford Shockley* (*1910*); *A. Philip Randolph* vs. *Asa Philip Randolph* (*1889*); *Norodom Sihanouk* vs. *King Norodom Sihanouk* (*1922*); and *Aaliyah* vs. *Aaliyah Haughton* (*1979*), among many other cases. Furthermore, the spelling of the name is sometimes different, e.g., *Mohammed Zahir Shah* vs. *Mohammad Zahir Shah* (*1914*).

Comparison to Previous Results

A set of extraction patterns relying on syntactically parsed text (e.g., *<subj> was kidnapped*) are acquired automatically from unannotated text in (Riloff 1996). After manual post-filtering, the patterns extract relations in the terrorism domain (perpetrator, victim, target of a terrorist event) from a set of 100 annotated documents, with an average precision of 58% and recall of 36%. A more sophisticated bootstrapping method (Riloff & Jones 1999) cautiously grows very small seed sets of five items, to less than 300 items after 50 consecutive iterations, with a final precision varying between 46% and 76% depending on the type of semantic lexicon. By adding the five best items extracted from 1700 text documents to the seed set after each iteration, 1000 semantic lexicon entries are collected after 200 iterations in (Thelen & Riloff 2002), at precision between 4.5% and 82.9%, again as a function of the target semantic type.

Changing the type of extracted items from semantic lexicons entries to binary relations (pairs of phrases), (Agichtein & Gravano 2000) exploits a collection of 300,000 news articles to iteratively expand a set of five seed relations of type Company-HeadquartersIn-Location. A key resource

[1] The assessment of precision depleted the resources that we were willing to spend on manual evaluation.

for identifying the components of candidate relations in text is a named entity recognizer that supports the Company and Location categories. The authors compute the precision manually for different recall levels, over a sample of 100 of the extracted relations. The precision varies from 93% for a recall of 20%, to 50% for a recall of 78%. A promising approach to extracting relations among named entities is introduced in (Hasegawa, Sekine, & Grishman 2004). A set of relations linking a Person with a GeopoliticalEntity, or a Company with another Company, are extracted from a collection containing news articles issued by a major newspaper over an entire year. For evaluation, the authors assemble manually a gold standard of approximately 250 items from the underlying collection. They report precision between 76% and 79%, and recall between 83% and 74%. The experiments described in (Lita & Carbonell 2004) consist in extracting up to 2000 new relations of various types including Person-Invents-Invention and Person-Founds-Organization, from a text collection of several gigabytes. However, the extracted relations are evaluated through their impact on a specific task, i.e., Question Answering, rather than through separate precision and recall metrics. In contrast to extracting relations from unstructured text, (Cucerzan & Agichtein 2005) derive shallow relations from the HTML tables in a collection 100 million Web documents. The resulting evaluation is also tied to the task of Question Answering, on which the authors indicate that the results are less than promising.

Several recent approaches specifically address the problem of extracting facts from Web documents. In (Cafarella *et al.* 2005), manually-prepared extraction rules are applied to a collection of 60 million Web documents to extract entities of types Company and Country, as well as facts of type Person-CeoOf-Company and City-CapitalOf-Country. Based on manual evaluation of precision and recall, a total of 23,128 company names are extracted at precision of 80%; the number decreases to 1,116 at precision of 90%. In addition, 2,402 Person-CeoOf-Company facts are extracted at precision 80%. The recall value is 80% at precision 90%. Recall is evaluated against the set of company names extracted by the system, rather than an external gold standard with pairs of a CEO and a company name. As such, the resulting metric for evaluating recall used in (Cafarella *et al.* 2005) is somewhat similar to, though more relaxed than, the MRR over the $\cap$Gold set described in the previous section.

Conclusion

Although orders of magnitude higher than previous results, the extraction of one million facts of a given type at approximately 90% precision is merely an intermediate checkpoint with respect to the broader goal of building large repositories of facts, as an aid in Web search. The next steps aim at increasing the number of extracted facts by another order of magnitude, while retaining similar precision levels, as well as performing experiments on other types of facts (including Language-SpokenIn-Country and Person-LeaderOf-Company). We are also exploring the role of generalized extraction patterns in automatically labeling and clustering the extracted facts.

References

Agichtein, E., and Gravano, L. 2000. Snowball: Extracting relations from large plaintext collections. In *Proceedings of the 5th ACM Conference on Digital Libraries (DL-00)*, 85–94.

Brants, T. 2000. TnT - a statistical part of speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP-00)*, 224–231.

Brin, S. 1998. Extracting patterns and relations from the world wide web. In *Proceedings of the 6th International Conference on Extending Database Technology (EDBT-98), Workshop on the Web and Databases*, 172–183.

Cafarella, M.; Downey, D.; Soderland, S.; and Etzioni, O. 2005. KnowItNow: Fast, scalable information extraction from the web. In *Proceedings of the Human Language Technology Conference (HLT-EMNLP-05)*, 563–570.

Collins, M., and Singer, Y. 1999. Unsupervised models for named entity classification. In *Proceedings of the 1999 Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)*, 189–196.

Cucerzan, S., and Agichtein, E. 2005. Factoid question answering over unstructured and structured content on the web. In *Proceedings of the 14th Text REtrieval Conference (TREC-05)*.

Dean, J., and Ghemawat, S. 2004. MapReduce: Simplified data processing on large clusters. In *Proceedings of the 6th Symposium on Operating Systems Design and Implementation (OSDI-04)*, 137–150.

Fleischman, M.; Hovy, E.; and Echihabi, A. 2003. Offline strategies for online question answering: Answering questions before they are asked. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, 1–7.

Hasegawa, T.; Sekine, S.; and Grishman, R. 2004. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 415–422.

Lin, D. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL-98)*, 768–774.

Lita, L., and Carbonell, J. 2004. Instance-based question answering: A data driven approach. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, 396–403.

Remy, M. 2002. Wikipedia: The free encyclopedia. *Online Information Review* 26(6):434.

Riloff, E., and Jones, R. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*, 474–479.

Riloff, E. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96)*, 1044–1049.

Thelen, M., and Riloff, E. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-02)*, 214–221.

Turney, P. 2001. Mining the web for synonyms: PMI-IR vs. LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning (ECML-01)*, 491–502.

Voorhees, E., and Tice, D. 2000. Building a question-answering test collection. In *Proceedings of the 23rd International Conference on Research and Development in Information Retrieval (SIGIR-00)*, 200–207.