Aligning Needles in a Haystack: Paraphrase Acquisition Across the Web

Marius Paşca and Péter Dienes

Google Inc.,
1600 Amphitheatre Parkway,
Mountain View, California, 94043, USA
{mars, dienes}@google.com

Abstract. This paper presents a lightweight method for unsupervised extraction of paraphrases from arbitrary textual Web documents. The method differs from previous approaches to paraphrase acquisition in that 1) it removes the assumptions on the quality of the input data, by using inherently noisy, unreliable Web documents rather than clean, trustworthy, properly formatted documents; and 2) it does not require any explicit clue indicating which documents are likely to encode parallel paraphrases, as they report on the same events or describe the same stories. Large sets of paraphrases are collected through exhaustive pairwise alignment of small needles, i.e., sentence fragments, across a haystack of Web document sentences. The paper describes experiments on a set of about one billion Web documents, and evaluates the extracted paraphrases in a natural-language Web search application.

1 Introduction

The information captured in textual documents frequently encodes semantically equivalent ideas through different lexicalizations. Indeed, given the generative power of natural language, different people employ different words or phrases to convey the same meaning, depending on factors such as background knowledge, level of expertise, style, verbosity and personal preferences. Two equivalent fragments of text may differ only slightly, as a word or a phrase in one of them is paraphrased in the other, e.g., through a synonym. Yet even small lexical variations represent challenges to any automatic decision on whether two text fragments have the same meaning, or are relevant to each other, since they are no longer lexically identical. Many natural-language intensive applications make such decisions internally. In document summarization, the generated summaries have a higher quality if redundant information has been discarded by detecting text fragments with the same meaning [1]. In information extraction, extraction templates will not be filled consistently whenever there is a mismatch in the trigger word or the applicable extraction pattern [2]. Similarly, a question answering system could incorrectly discard a relevant document passage based on the absence of a question phrase deemed as very important [3], even if the passage actually contains a legitimate paraphrase.

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 119–130, 2005.

In information retrieval, deciding whether a text fragment (e.g., a document) is relevant to another text fragment (i.e., the query) is crucial to the overall output, rather than merely useful within some internal system module. Indeed, relevant documents or passages may be missed, due to the apparent mismatch between their terms and the paraphrases occurring in the users' queries. The previously proposed solutions to the mismatch problem vary with respect to the source of the data used for enriching the query with alternative terms. In automatic query expansion, the top documents provide additional query terms [4]. An alternative is to attempt to identify the concepts captured in the queries and find semantically similar concepts in external resources, e.g., lexical databases [5, 6]. This paper explores a different direction, namely the unsupervised acquisition of large sets of paraphrases from unstructured text within Web documents, and their exploitation in natural-language Web search.

We present a lightweight method for unsupervised extraction of paraphrases from arbitrary, textual Web documents. The method taps the textual contents provided by millions of anonymous Web document contributors. The remainder of the paper is structured as follows. After a condensed overview of the paraphrase acquisition method and a contrast to previous literature in Section 2, Section 3 presents the method in more detail. Section 4 describes evaluation results when applying the method to textual documents from a Web repository snapshot of the Google search engine.

2 Method at a Glance

The proposed acquisition method collects large sets of word and phrase-level paraphrases via exhaustive pairwise alignment of small needles, i.e., sentence fragments, across a haystack of Web document sentences. The acquisition of paraphrases is a side-effect of the alignment.

In the example in Figure 1, if two sentence fragments have common word sequences at both extremities, then the variable word sequences in the middle are potential paraphrases of each other. A significant advantage of this extraction mechanism is that it can acquire paraphrases from sentences whose information content overlaps only partially, as long as the fragments align. Indeed, the source sentences of the paraphrase (*withdrew from*, *pulled out of*), as well as of (*took effect*, *came into force*), are arguably quite different overall in Figure 1. Moreover, the sentences are part of documents whose content intersection is very small.

In addition to its relative simplicity when compared to more complex, sentence-level paraphrase acquisition [7], the method introduced in this paper is a departure from previous approaches in several respects. First, the paraphrases are not limited to variations of specialized, domain-specific terms as in [8], nor are they restricted to a narrow class such as verb paraphrases [9]. Second, as opposed to virtually all previous approaches, the method does not require high-quality, clean, trustworthy, properly-formatted input data. Instead, it uses inherently noisy, unreliable Web documents. The source data in [10] is also a set of Web documents. However, it is based on top search results collected

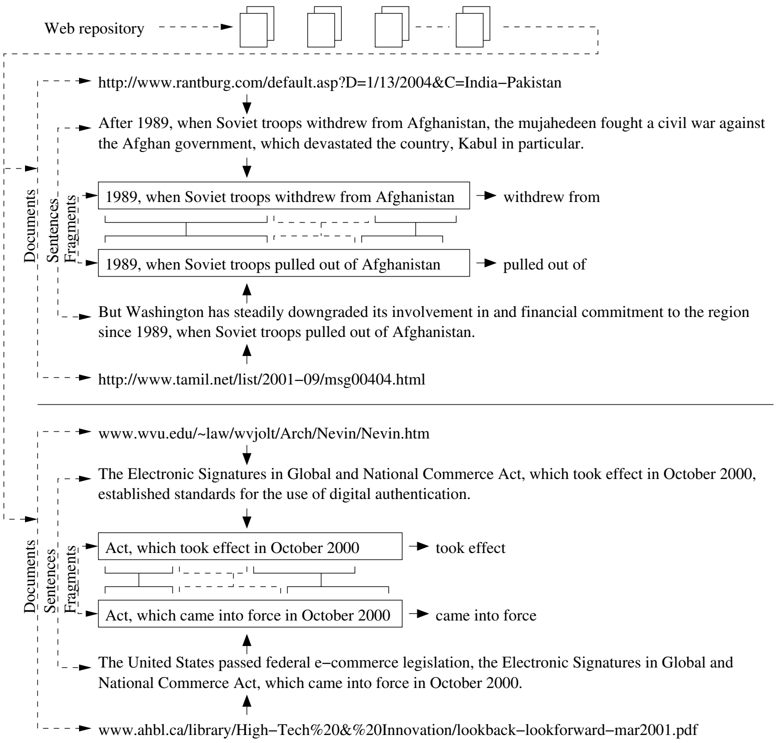
Web repository
http://www.rantburg.com/default.asp?D=1/13/2004&C=India–Pakistan
After 1989, when Soviet troops withdrew from Afghanistan, the mujahedeen fought a civil war against the Afghan government, which devastated the country, Kabul in particular.
1989, when Soviet troops withdrew from Afghanistan
withdrew from
1989, when Soviet troops pulled out of Afghanistan
pulled out of
But Washington has steadily downgraded its involvement in and financial commitment to the region since 1989, when Soviet troops pulled out of Afghanistan.
http://www.tamil.net/list/2001–09/msg00404.html
Documents
Sentences
Fragments
www.wvu.edu/~law/wvjolt/Arch/Nevin/Nevin.htm
The Electronic Signatures in Global and National Commerce Act, which took effect in October 2000, established standards for the use of digital authentication.
Act, which took effect in October 2000
took effect
Act, which came into force in October 2000
came into force
The United States passed federal e–commerce legislation, the Electronic Signatures in Global and National Commerce Act, which came into force in October 2000.
www.ahbl.ca/library/High–Tech%20&%20Innovation/lookback–lookforward–mar2001.pdf
Documents
Sentences
Fragments

Fig. 1. Paraphrase acquisition from unstructured text across the Web

from external search engines, and its quality benefits implicitly from the ranking functions of the search engines. Third, the input documents here are not restricted to a particular genre, whereas virtually all other recent approaches are designed for collections of parallel news articles, whether the articles are part of a carefully-compiled collection [11] or aggressively collected from Web news sources [12]. Fourth, the acquisition of paraphrases in this paper does not rely on external clues and attributes that two documents are parallel and must report on the same or very similar events. Comparatively, previous work has explicit access to, and relies strongly on clues such as the same or very similar timestamps being associated to two news article documents [11], or knowledge that two documents are translations by different people of the same book into the same language [13].

3 Mining the Web for Paraphrases

The use of the Web as input data source strongly impacts the design of the method, since the average Web document is much noisier and less reliable than documents in standard textual collections. Furthermore, the separation of useful textual information from other items within the document is trivial in standard collections. In contrast, Web documents contain extraneous HTML information, formatting errors, intra- and inter-document inconsistencies, spam and other adversarial information, and in general they lack any assumptions regarding a common document structure. Consequently, the acquisition of paraphrases must be robust, handle Web documents with only minimal linguistic processing, avoid expensive operations, and scale to billions of sentences.

3.1 Document Pre-processing

As a pre-requisite to the actual acquisition of paraphrases, the Web documents are converted from raw string representations into more meaningful linguistic units. After filtering out HTML tags, the documents are tokenized, split into sentences and part-of-speech tagged with the TnT tagger [14]. Many of the candidate sentences are in fact random noise caused by the inconsistent structure (or complete lack thereof) of Web documents, among other factors. To improve the quality of the data, sentences are retained for further processing only if they satisfy the following lightweight sanity checks: 1) they are reasonably sized: sentences containing less than 5 words or more than 30 words are discarded; 2) they contain at least one verb that is neither a gerund nor a modal verb; 3) they contain at least one non-verbal word starting in lower-case; 4) none of the words is longer than 30 characters; and 5) less than half of the words are numbers. Since the experiments use a collection of English documents, these checks are geared towards English.

3.2 Acquisition via Text Fragment Alignment

At Web scale, the number of sentences that pass the fairly aggressive sanity checks during document pre-processing is still extremely large, easily exceeding one billion. Any brute-force alignment of all pairs of document sentences is therefore unfeasible. Instead, the acquisition of paraphrases operates at the level of text fragments (ngrams) as shown in Figure 2.

The extraction algorithm roughly consists of the following three phases:

- Generate candidate ngrams from all sentences (steps 1 through 5 in Figure 2);
- Convert each ngram into a ready-to-align pair of a variable fragment (a candidate paraphrase) and a constant textual anchor (steps 6 through 13);
- Group the pairs with the same anchors; collect the variable fragments within each group of pairs as potential paraphrases of one another (steps 14 to 20).

The algorithm starts with the generation of candidate ngrams, by collecting all possible ngrams such that their length varies within pre-defined boundaries. More precisely, an ngram starts and ends in a fixed number of words (L_C); the count of the additional (ngram) words in-between varies within pre-defined limits (Min_P and Max_P, respectively).

The concatenation of the fixed-length left (Cst_L) and right (Cst_R) extremities of the ngram forms a textual **anchor** for the variable fragment (Var) in the middle. The variable fragment becomes a potential candidate for a paraphrase:

```
Input:
  ${S}$ set of sentences
  $L_C$ length of constant extremities
  $Min_P$, $Max_P$ paraphrase length bounds
Vars:
  ${N}$ set of ngrams with attached info
  ${P}$ set of pairs (anchor, candidate)
  ${R}$ set of paraphrase pairs with freq info
Output: ${R}$
Steps:
1 ${R} = {N} = {P}$ = empty set;
2 For each sentence $S_i$ in ${S}$
3   Generate ngrams $N_{ij}$ between length
    $2 \times L_C + Min_P$ and $2 \times L_C + Max_P$
4   For each $N_{ij}$, attach addtl. info $Att_{ij}$
5     Insert $N_{ij}$ with $Att_{ij}$ into ${N}$
6 For each ngram $N_i$ in ${N}$
7   $L_{N_i}$ = length of $N_i$
8   $Cst_{L|}$ = subseq [0, $L_C$-1] of $N_i$
9   $Cst_R$ = subseq [$L_{N_i}L_C$, $L_{N_i}$-1] of $N_i$
10  $Var_i$ = subseq [$L_C$, $L_{N_i}$-$L_C$-1] of $N_i$
11  $Anchor_i$ = concat of $Cst_{L|}$ and $Cst_R$
12  $Anchor_i$ = concat of $Att_i$ and $Anchor_i$
13  Insert pair ($Anchor_i$,$Var_i$) into ${P}$
14 Sort pairs in ${P}$ based on their anchor
15 For each ${P_i} \subset {P}$ with same anchor
16   For all item pairs $P_{i_1}$ and $P_{i_2}$ in ${P_i}$
17     $Var_{i_1}$ = variable part of pair $P_{i_1}$
18     $Var_{i_2}$ = variable part of pair $P_{i_2}$
19     Incr. count of ($Var_{i_1}$,$Var_{i_2}$) in ${R}$
20     Incr. count of ($Var_{i_2}$,$Var_{i_1}$) in ${R}$
21 Return ${R}$
```

Fig. 2. Algorithm for paraphrase acquisition from Web document sentences

(S$_1$) *Together they form the Platte River*, $\underbrace{\textit{which eventually}}_{Cst_L}$ $\underbrace{\textit{flows}}_{Var}$ $\underbrace{\textit{into the Gulf}}_{Cst_R}$ *of Mexico.*

Whenever the anchors of two or more ngrams are the same, their variable fragments are considered to be potential paraphrases of each other, thus implementing a const-var-const type of alignment.

3.3 Alignment Anchors

According to the simplified discussion from above, the algorithm in Figure 2 may align two sentence fragments *"decided to read the government report published last month"* and *"decided to read the edition published last month"* to incorrectly produce *government report* and *edition* as potential paraphrases of each other. To avoid such alignments, Steps 4 and 12 of the algorithm enrich the anchoring text around each paraphrase candidate, namely by extending the anchors to include additional information from the source sentence. By doing so, the anchors become longer and more specific, and thus closer to expressing the same information content. In turn, this reduces the chances of any two ngrams to align, since ngram alignment requires the complete matching of the corresponding anchors. In other words, the amount of information captured in the anchors is a trade-off between coverage (when anchors are less specific) and accuracy of the acquired paraphrases (when the anchors are more specific). At the low end, less specific anchors include only immediate contextual information. This corresponds to the algorithm in Figure 2, when nothing is attached to any of the ngrams in Step 4. At the high end, one could collect all the remaining words of the sentence outside the ngram, and attach them to more specific anchors in Step 4. This is equivalent to pairwise alignment of full-length sentences.

We explore three different ways of collecting additional anchoring information from the sentences:

Table 1. Examples of paraphrase pairs collected from the Web with one of Ngram-Entity or Ngram-Relative, but not with the other

Only with Ngram-Entity	Only with Ngram-Relative
abduction, kidnapping	abolished, outlawed
bachelor degree, bachelors degree	abolished slavery, freed the slaves
cause, result in	causes, results in
indicate, specify	carries, transmits
inner product space, vector space	died from, succumbed to
kill, murder	empties into, flows to
obligations, responsibilities	funds, pays for
registered service marks, registered trademarks	means, stands for
video poker betting, video poker gambling	penned, wrote
x-mas gift, x-mas present	seized, took over

– Ngram-Only: The anchor includes only the contextual information assembled from the fixed-length extremities of the ngram. Nothing else is attached to the anchor.
– Ngram-Entity: In addition to Ngram-Only, the anchor contains the preceding and following named entities that are nearest to the ngram. Sentences without such named entities are discarded. The intuition is that the ngram contains information which relates the two entities to each other.
– Ngram-Relative: On top of Ngram-Only, the anchor includes the remaining words of the adverbial relative clause in which the variable part of the ngram appears, e.g., *"when Soviet Union troops pulled out of Afghanistan"*, or *"which came into force in 2000"* in Figure 1. The clause must modify a named entity or a date, which is also included in the anchor. Sentences not containing such clauses are rejected. [1] The intuitive motivation in that the entity is related to part of the ngram via the adverbial particle.

For illustration, consider the earlier example of the sentence S_1 from Section 3.2. With Ngram-Entity, *Platte River* (preceding entity) and *Mexico* (following entity) are included in the anchor. In comparison, with Ngram-Relative the additional information combines *Platte River* (entity) and *of Mexico* (remainder of relative clause). In this example, the difference between Ngram-Entity and Ngram-Relative happens to be quite small. In general, however, the differences are more significant. Table 1 illustrates paraphrases collected from the Web by only one of the two anchoring mechanisms.

To ensure robustness on Web document sentences, simple heuristics rather than complex tools are used to approximate the additional information attached to ngrams in Ngram-Entity and Ngram-Relative. Named entities are approximated by proper nouns, as indicated by part-of-speech tags. Adverbial relative clauses, together with the entities or dates they modify, are detected according to a small set of lexico-syntactic patterns which can be summarized as:

$\langle [\mathit{Date}|\mathit{Entity}]\ [,|\text{-}|(|\text{nil}]\ [\mathit{Wh}]\ \mathit{RelClause}\ [,|\text{-}|)|.]\rangle$

[1] By discarding many sentences, Ngram-Relative sacrifices recall in favor of precision.

where *Wh* is one of *who*, *when*, *which* or *where*. The patterns are based mainly on *wh*-words and punctuation. The matching adverbial clause *RelClause* must satisfy a few other constraints, which aim at avoiding, rather than solving, complex linguistic phenomena. First, personal and possessive pronouns are often references to other entities. Therefore clauses containing such pronouns are discarded as ambiguous. Second, appositives and other similar pieces of information are confusing when detecting the end of the current clause. Consequently, during pattern matching, if the current clause does not contain a verb, the clause is either extended to the right, or discarded upon reaching the end of the sentence.

4 Evaluation

The input data for paraphrase acquisition is a collection of 972 million Web documents, from a Web repository snapshot of the Google search engine taken in 2003. All documents are in English. The parameters controlling the length of the ngrams and candidate paraphrases, introduced in Figure 2, are L_C=3, Min_P=1 and Max_P=4. [2] The anchors use additional information from the sentences, resulting in separate runs and sets of paraphrases extracted with Ngram-Only, Ngram-Entity and Ngram-Relative respectively. The experiments use a parallel programming model [15]. The extracted paraphrase pairs that co-occur very infrequently (i.e., in less than 5 unique ngram pairs) are discarded.

4.1 Quantitative Results

The sanity checks applied in document pre-processing (see Section 3.1) discard a total of 187 billion candidate sentences from the input documents, with an average of 3 words per sentence. In the case of Ngram-Only, paraphrases are extracted from the remaining 9.5 billion sentences, which have 17 words on average. As explained in Section 3.3, Ngram-Entity and Ngram-Relative apply a set of additional constraints as they search the sentences for more anchoring information. Ngram-Entity discards 72 million additional sentences. In contrast, as many as 9.3 billion sentences are rejected by the constraints encoded in Ngram-Relative.

The number of paraphrase pairs extracted from the Web varies with the particular kind of anchoring mechanism. The simplest one, i.e., Ngram-Only, produces 41,763,994 unique pairs that co-occur in at least 5 different ngrams. With Ngram-Relative, the output consists of 13,930 unique pairs. In comparison, Ngram-Entity generates 101,040 unique pairs. Figure 3 shows that the number of acquired paraphrases varies more or less linearly in the size of the input data.

The large majority of the paraphrase pairs contain either two single-word phrases (40% for Ngram-Entity, and 49% for Ngram-Relative), or one single-word and one multi-word phrase (22% for Ngram-Entity, and 43% for Ngram-Relative). Table 2 illustrates the top paraphrase pairs with two multi-word phrases, after removal of paraphrases containing only stop words, or upper/lower

[2] No experiments were performed with higher values for Max_P (to collect longer paraphrases), or higher/lower values for L_C (to use more/less context for alignment).

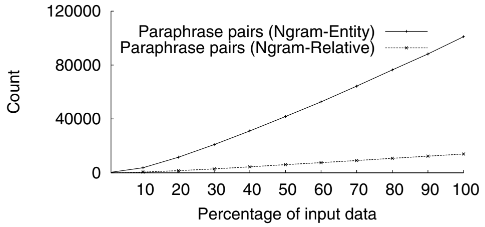

Fig. 3. Variation of the number of acquired paraphrase pairs with the input data size

Table 2. Top ranked multi-word paraphrase pairs in decreasing order of frequency of co-occurrence

#	Ngram-Entity	Ngram-Relative
1	DVD Movie, VHS Movie	became effective, took effect
2	betting is excited, wagering is excited	came into force, took effect
3	betting is, wagering is	became effective, went into effect
4	betting is excited, gambling is excited	became effective, came into force
5	Annual Meeting of, meeting of	became effective, came into effect
6	center of, centre of	entered into force, took effect
7	betting is, gambling is	one hour, two hours

case variation. Top multi-word phrases extracted by Ngram-Relative tend to be self-contained syntactic units. For instance, *entered into force* is a verb phrase in Table 2. In contrast, many of the top paraphrases with Ngram-Entity end in a linking word, such as the pair (*center of*, *centre of*). Note that every time this pair is extracted, the smaller single-word paraphrase pair that folds the common linking word into the anchor, e.g., (*center*, *centre*), is also extracted.

4.2 Quality of Paraphrases

Table 2 shows that the extracted paraphrases are not equally useful. The pair (*became effective*, *took effect*) is arguably more useful than (*one hour*, *two hours*). Table 3 is a side-by-side comparison of the accuracy of the paraphrases with Ngram-Only, Ngram-Entity and Ngram-Relative respectively. The values are the result of manual classification of the top, middle and bottom 100 paraphrase pairs from each run into 11 categories. The first six categories correspond to pairs classified as correct. For instance (*Univeristy*, *University*) is classified in class (1); (*Treasury*, *treasury*) in (2); (*is*, *are*) in (3); (*e-mail*, *email*) in (4); and (*can*, *could*) in (5). The pairs in class (6) are considered to be the most useful; they include (*trip*, *visit*), (*condition*, *status*), etc. The next three classes do not contain synonyms but are still useful. The pairs in (7) are siblings rather than direct synonyms; examples are (*twice a year*, *weekly*) and (*French*, *welsh*). Furthermore, modal verbs such as (*may*, *should*), numbers, and prepositions like (*up*, *back*) also fall under class (7). Many of the 63 pairs classified as siblings

Table 3. Quality of the acquired paraphrases

	Ngram-Only			Ngram-Entity			Ngram-Relative		
Classification of Pairs	Top	Mid	Low	Top	Mid	Low	Top	Mid	Low
	100	100	100	100	100	100	100	100	100
(1) Correct; punct., symbols, spelling	1	5	11	12	6	20	18	11	15
(2) Correct; equal if case-insensitive	0	5	0	27	2	11	9	2	14
(3) Correct; both are stop words	4	0	0	3	0	1	1	0	0
(4) Correct; hyphenation	0	1	4	10	35	8	2	19	43
(5) Correct; morphological variation	8	1	10	9	10	20	20	15	6
(6) Correct; synonyms	16	8	21	5	32	14	33	23	6
Total correct	29	20	46	66	85	74	83	70	84
(7) Siblings rather than synonyms	63	29	19	32	8	15	5	7	7
(8) One side adds an elaboration	0	0	3	0	0	0	1	2	1
(9) Entailment	0	3	2	0	0	1	3	1	0
Total siblings	63	32	24	32	8	16	9	10	8
(10) Incorrect; antonyms	6	0	2	0	1	4	4	3	4
(11) Incorrect; other	2	48	28	2	6	6	4	17	4
Total incorrect	8	48	30	2	7	10	8	20	8

with Ngram-Only in Table 3 are precisely such words. Class (8) contains pairs in which a portion of one of the elements is a synonym or phrasal equivalent of the other element, such as (*poliomyelitis globally*, *polio*) and (*UNC*, *UNC-CH*), whereas (9) captures what can be thought of as entailment, e.g., (*governs*, *owns*) and (*holds*, *won*). Finally, the last two classes from Table 3 correspond to incorrect extractions, due to either antonyms like (*lost*, *won*) and (*your greatest strength*, *your greatest weakness*) in class (10), or other factors in (11).

The aggregated evaluation results, shown in bold in Table 3, suggest that Ngram-Only leads to paraphrases of lower quality than those extracted with Ngram-Entity and Ngram-Relative. In particular, the samples from the middle and bottom of the Ngram-Only paraphrases contain a much higher percentage of incorrect pairs. The results also show that, for Ngram-Entity and Ngram-Relative, the quality of paraphrases is similar at different ranks in the paraphrase lists sorted by the number of different ngrams they co-occur in. For instance, the total number of correct pairs has comparable values for the top, middle and bottom pairs. This confirms the usefulness of the heuristics introduced in Section 3.3 to discard irrelevant sentences with Ngram-Entity and Ngram-Relative.

4.3 Paraphrases in Natural-Language Web Search

The usefulness of paraphrases in Web search is assessed via an existing experimental repository of more than 8 million factual nuggets associated with a date. Repositories of factual nuggets are built offline, by matching lightweight, open-domain lexico-semantic patterns on unstructured text. In the repository used in this paper, a factual nugget is a sentence fragment from a Web document, paired with a date extracted from the same document, when the event encoded in the

Table 4. Impact of expansion of the test queries (QH/QL=count of queries with higher/lower scores than without expansion, NE=Ngram-Entity, NR=Ngram-Relative)

Max. nr. disjunctions	QH		QL		Score	
per expanded phrase	NE	NR	NE	NR	NE	NR
1 (no paraphrases)	0	0	0	0	52.70	52.70
2 (1 paraphrase)	17	8	7	6	64.50	57.62
3 (2 paraphrases)	22	13	6	9	70.38	60.46
4 (3 paraphrases)	23	15	6	7	71.42	60.39
5 (4 paraphrases)	26	18	12	5	71.73	63.35

sentence fragment occurred according to the text, e.g., ⟨*1937, Golden Gate was built*⟩, and ⟨*1947, Bell Labs invented the transistor*⟩.

A test set of temporal queries is used to extract direct results (dates) from the repository of factual nuggets, by matching the queries against the sentence fragments, and retrieving the associated dates. The test queries are all queries that start with either *When* or *What year*, namely 207 out of the total count of 1893 main-task queries, from the Question Answering track [16] of past editions (1999 through 2002). The metric for measuring the accuracy of the retrieved results is the de-facto scoring metric for fact-seeking queries, that is, the reciprocal rank of the first returned result that is correct (in the gold standard) [16]. If there is no correct result among the top 10 returned, the query receives no credit. Individual scores are aggregated (i.e., summed) over the entire query set.

In a series of parallel experiments, all phrases from the test queries are expanded into Boolean disjunctions with their top-ranked paraphrases. Query words with no paraphrase are placed into the expanded queries in their original form. The other query words are expanded only if they are single words, for simplicity. Examples of implicitly-Boolean queries expanded disjunctively, before removal of stop words and *wh*-words, are:

– When did Amtrak (begin | start | began | continue | commence) (operations | operation | activities | Business | operational)?
– When was the De Beers (company | Co. | firm | Corporation | group) (founded | established | started | created | co-founded)?

Table 4 illustrates the impact of paraphrases on the accuracy of the dates retrieved from the repository of factual nuggets associated with dates. When compared to non-expanded queries, paraphrases consistently improve the accuracy of the returned dates. Incremental addition of more paraphrases results in more individual queries with a better score than for their non-expanded version, and higher overall scores for the returned dates. The paraphrases extracted with Ngram-Entity produce scores that are higher than those of Ngram-Relative, due mainly to higher coverage. Since the temporal queries represent an external, objective test set, they provide additional evidence regarding the quality of paraphrases in a practical application.

5 Conclusion

The Web has gradually grown into a noisy, unreliable, yet powerful resource of human knowledge. This knowledge ranges from basic word usage statistics to intricate facts, background knowledge and associated inferences made by humans reading Web documents. This paper describes a method for unsupervised acquisition of lexical knowledge across the Web, by exploiting the numerous textual forms that people use to share similar ideas, or refer to common events. Large sets of paraphrases are collected through pairwise alignment of ngrams occurring within the unstructured text of Web documents. Several mechanisms are explored to cope with the inherent lack of quality of Web content. The quality of the extracted paraphrases improves significantly when the textual anchors used for aligning potential paraphrases attempt to approximate, even at a very coarse level, the presence of additional information within the sentences. In addition to the known role of the extracted paraphrases in natural-language intensive applications, the experiments in this paper illustrate their impact in returning direct results to natural-language queries.

The final output of the extraction algorithm lacks any distinction among paraphrases that apply to only one of the several senses or part of speech tags that a word or phrase may have. For instance, *hearts*, *center* and *middle* mix the medical and positioning senses of the word heart. Conversely, the extracted paraphrases may capture only one sense of the word, which may not match the sense of the same word in the queries. As an example, in the expansion of one of the test queries, *"Where is the massive North Korean (nuclear|atomic) (complex|real) (located|situated|found)?"*, a less-than-optimal paraphrase of *complex* not only provides a sibling rather than a near synonym, but may incorrectly shift the focus of the search towards the mathematical sense of the word (*complex* versus *real* numbers). Aggregated contextual information from the source ngrams could provide a means for selecting only some of the paraphrases, based on the query. As another direction for future work, we plan to revise the need for language-dependent resources (namely, the part of speech tagger) in the current approach, and explore possibilities of minimizing or removing their use for seamless transfer of the approach to other languages.

References

1. Hirao, T., Fukusima, T., Okumura, M., Nobata, C., Nanba, H.: Corpus and evaluation measures for multiple document summarization with multiple sources. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING-04), Geneva, Switzerland (2004) 535–541
2. Shinyama, Y., Sekine, S.: Paraphrase acquisition for information extraction. In: Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL-03), 2nd Workshop on Paraphrasing: Paraphrase Acquisition and Applications, Sapporo, Japan (2003) 65–71
3. Paşca, M.: Open-Domain Question Answering from Large Text Collections. CSLI Studies in Computational Linguistics. CSLI Publications, Distributed by the University of Chicago Press, Stanford, California (2003)

4. Mitra, M., Singhal, A., Buckley, C.: Improving automatic query expansion. In: Proceedings of the 21st ACM Conference on Research and Development in Information Retrieval (SIGIR-98), Melbourne, Australia (1998) 206–214
5. Schutze, H., Pedersen, J.: Information retrieval based on word senses. In: Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval. (1995) 161–175
6. Zukerman, I., Raskutti, B.: Lexical query paraphrasing for document retrieval. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING-02), Taipei, Taiwan (2002) 1177–1183
7. Barzilay, R., Lee, L.: Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In: Proceedings of the 2003 Human Language Technology Conference (HLT-NAACL-03), Edmonton, Canada (2003) 16–23
8. Jacquemin, C., Klavans, J., Tzoukermann, E.: Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In: Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics (ACL-97), Madrid, Spain (1997) 24–31
9. Glickman, O., Dagan, I.: Acquiring Lexical Paraphrases from a Single Corpus. In: Recent Advances in Natural Language Processing III. John Benjamins Publishing, Amsterdam, Netherlands (2004) 81–90
10. Duclaye, F., Yvon, F., Collin, O.: Using the Web as a linguistic resource for learning reformulations automatically. In: Proceedings of the 3rd Conference on Language Resources and Evaluation (LREC-02), Las Palmas, Spain (2002) 390–396
11. Shinyama, Y., Sekine, S., Sudo, K., Grishman, R.: Automatic paraphrase acquisition from news articles. In: Proceedings of the Human Language Technology Conference (HLT-02), San Diego, California (2002) 40–46
12. Dolan, W., Quirk, C., Brockett, C.: Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING-04), Geneva, Switzerland (2004) 350–356
13. Barzilay, R., McKeown, K.: Extracting paraphrases from a parallel corpus. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-01), Toulouse, France (2001) 50–57
14. Brants, T.: TnT - a statistical part of speech tagger. In: Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP-00), Seattle, Washington (2000) 224–231
15. Dean, J., Ghemawat, S.: MapReduce: Simplified data processing on large clusters. In: Proceedings of the 6th Symposium on Operating Systems Design and Implementation (OSID-04), San Francisco, California (2004) 137–150
16. Voorhees, E., Tice, D.: Building a question-answering test collection. In: Proceedings of the 23rd International Conference on Research and Development in Information Retrieval (SIGIR-00), Athens, Greece (2000) 200–207