Mining Paraphrases from Self-anchored Web Sentence Fragments

Marius Paşca

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

Abstract. Near-synonyms or paraphrases are beneficial in a variety of natural language and information retrieval applications, but so far their acquisition has been confined to clean, trustworthy collections of documents with explicit external attributes. When such attributes are available, such as similar time stamps associated to a pair of news articles, previous approaches rely on them as signals of potentially high content overlap between the articles, often embodied in sentences that are only slight, paraphrase-based variations of each other. This paper introduces a new unsupervised method for extracting paraphrases from an information source of completely different nature and scale, namely unstructured text across arbitrary Web textual documents. In this case, no useful external attributes are consistently available for all documents. Instead, the paper introduces linguistically-motivated text anchors, which are identified automatically within the documents. The anchors are instrumental in the derivation of paraphrases through lightweight pairwise alignment of Web sentence fragments. A large set of categorized names, acquired separately from Web documents, serves as a filtering mechanism for improving the quality of the paraphrases. A set of paraphrases extracted from about a billion Web documents is evaluated both manually and through its impact on a natural-language Web search application.

1 Motivation

The qualitative performance of applications relying on natural language processing may suffer, whenever the input documents contain text fragments that are lexically different and yet semantically equivalent as they are paraphrases of each other. The automatic detection of paraphrases is important in document summarization, to improve the quality of the generated summaries [1]; information extraction, to alleviate the mismatch in the trigger word or the applicable extraction pattern [2]; and question answering, to prevent a relevant document passage from being discarded due to the inability to match a question phrase deemed as very important [3].

In specialized collections such as news, the coverage of major events by distinct sources generates large numbers of documents with high overlap in their content. Thus, the task of detecting documents containing similar or equivalent

A. Jorge et al. (Eds.): PKDD 2005, LNAI 3721, pp. 193–204, 2005.

information is somewhat simplified by the relative document homogeneity, use of properly-formatted text, the availability of external attributes (headlines), and knowledge of the document temporal proximity (similar article issue dates, or time stamps). When switching to unrestricted Web textual documents, all these advantages and clues are lost. Yet despite the diversity of content, the sheer size of the Web suggests that text fragments "hidden" inside quasi-anonymous documents will sometimes contain similar, or even equivalent information.

The remainder of the paper is structured as follows. After an overview of the proposed paraphrase acquisition method and a contrast to previous literature in Section 2, Section 3 provides more details and explains the need for self-anchored fragments as a source of paraphrases, as well as extensions for increasing the accuracy. Candidate paraphrases are filtered based on a large set of categorized named entities acquired separately from unstructured text. Section 4 describes evaluation results when applying the method to textual documents from a Web repository snapshot of the Google search engine. The section also evaluates the impact of the extracted paraphrases in providing search results that directly answer a standard evaluation set of natural-language questions.

2 Proposed Method for Paraphrase Acquisition

2.1 Goals

With large content providers and anonymous users contributing to the information accessible online, the Web has grown into a significant resource of implicitly-encoded human knowledge. The lightweight unsupervised method, presented in this paper, acquires useful paraphrases by mining arbitrary textual documents on the Web. The method is designed with a few goals in mind, which also represent advantages over previous methods:

1. No assumptions of any kind are made about the source, genre or structure of the input documents. In the experiments reported here, noise factors such as errors, misspellings, improperly formed sentences, or the use of HTML tags as implicit visual delimiters of sentences, are the norm rather than exceptions.
2. The method does not have access to any document-level attributes, which might otherwise hint at which pairs of documents are more likely to be sources of paraphrases. Such external attributes are simply not available for Web documents.
3. The acquisition is lightweight, robust and applicable to Web-scale collections. This rules out the use of deeper text analysis tools, e.g. syntactic [4] or semantic-role parsers [5].
4. For simplicity, the method derives paraphrases as a by-product of pairwise alignment of sentence fragments. When the extremities of the sentence fragments align, the variable parts become potential paraphrases of each other.
5. The method places an emphasis on defining the granularity (e.g., words, phrases, sentences or entire passages) and the actual mechanism for selecting the sentence fragments that are candidates for pairwise alignment. The

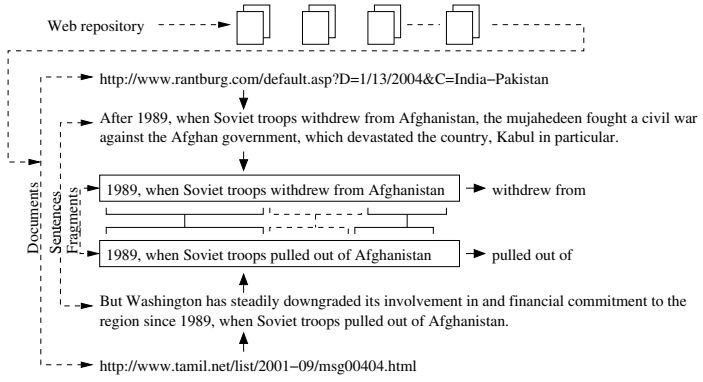

Fig. 1. Paraphrase acquisition from unstructured text across the Web

selection depends on *text anchors*, which are linguistic patterns whose role is twofold. First, they reduce the search/alignment space, which would otherwise be overwhelming (i.e., all combinations of contiguous sentence fragments). Second and more important, the anchors increase the quality of potential paraphrases, as they provide valuable linguistic context to the alignment phase, with little processing overhead.

2.2 Overview of Acquisition Method

As a pre-requisite, after filtering out HTML tags, the documents are tokenized, split into sentences and part-of-speech tagged with the TnT tagger [6]. Due to the inconsistent structure (or complete lack thereof) of Web documents, the resulting candidate sentences are quite noisy. Therefore some of the burden of identifying reliable sentences as sources of paraphrases is passed onto the acquisition mechanism.

Figure 1 illustrates the proposed method for unsupervised acquisition of paraphrases from Web documents. To achieve the goals listed above, the method mines Web documents for sentence fragments and associated text anchors. The method consists in searching for pairwise alignments of text fragments that have the same associated anchors. In the example, the anchors are identical time stamps (i.e., *1989*) of the events captured by the sentence fragments. The acquisition of paraphrases is a side-effect of the alignment.

The choice of the alignment type determines the constraints that two sentence fragments must satisfy in order to align, as well as the type of the acquired paraphrases. The example in Figure 1 is a const-var-const alignment. The two sentence fragments must have common word sequences at both extremities, as well as identical associated anchors. If that constraint holds, then the middle, variable word sequences are potential paraphrases of each other. Even if two sentences have little information content in common, their partial overlap can still produce paraphrase pairs such as ⟨*pulled out of*, *withdrew from*⟩ in Figure 1.

2.3 Comparison to Previous Work

Lexical resources such as WordNet [7] offer access to synonym sets, at the expense of many years of manual construction efforts. As general-purpose resources, they only cover the upper ontologies of a given language. Misspellings, idioms and other non-standard linguistic phenomena occur often in the noisy Web, but are not captured in resources like WordNet. Search engine hit counts rather than entries in lexical resources can be successfully exploited to select the best synonym of a given word, out of a small, closed-set of possible synonyms [8].

In addition to its relative simplicity when compared to more complex, sentence-level paraphrase acquisition [9], the method introduced in this paper is a departure from previous data-driven approaches in several respects. First, the paraphrases are not limited to variations of specialized, domain-specific terms as in [10], nor are they restricted to a narrow class such as verb paraphrases [11]. 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 [12] is also a set of Web documents. However, it is based on top search results collected 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 [13] or aggressively collected from Web news sources [14]. 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 [13], or knowledge that two documents are translations by different people of the same book into the same language [15].

3 Anchored Sentence Fragments as Sources of Paraphrases

Even though long-range phrase dependencies often occur within natural-language sentences, such dependencies are not available without deeper linguistic processing. Therefore the acquisition method exploits only short-range dependencies, as captured by text fragments that are *contiguous* sequences of words. Two factors contribute substantially to the quality of the extracted paraphrases, namely the granularity of the text fragments, and the selection of their boundaries.

3.1 Fragment Granularity: Passages vs. Sentence Fragments

In principle, the granularity of the text fragments used for alignment ranges from full passages, a few sentences or a sentence, down to a sentence fragment, a phrase

Table 1. Examples of incorrect paraphrase pairs collected through the alignment of sentence fragments with arbitrary boundaries

(Wrong) Pairs	Examples of Common Sentence Fragments	
⟨*city*, *place*⟩	(to visit the _ of their birth)	(is a beautiful _ on the river),
	(live in a _ where things are)	(once the richest _ in the world)
⟨*dogs*, *men*⟩	(one of the _ took a step)	(does not allow _ to live in),
	(average age of _ at diagnosis is)	(a number of _ killed and wounded)

or a word. In practice, full text passages provide too much alignment context to be useful, as the chance of finding pairwise const-var-const alignments of any two passage pairs is very low. On the other hand, words and even phrases are useless since they are too short and do not provide any context for alignment. Sentences offer a good compromise in terms of granularity, but they are rarely confined to describing exactly one event or property as illustrated by the two sentences from Figure 1. Even though both sentences use similar word sequences to refer to a common event, i.e. the withdrawal of troops, they do not align to each other as complete sequences of words. Based on this and other similar examples, the paraphrase acquisition method relies on the alignment of contiguous chunks of sentences, or *sentence fragments*, instead of full-length sentences.

3.2 Fragment Boundaries: Arbitrary vs. Self-Anchored

It is computationally impractical to consider all possible sentence fragments as candidates for alignment. More interestingly, such an attempt would actually strongly degrade the quality of potential extractions as shown in Table 1. The pairs ⟨*city*, *place*⟩ and ⟨*dogs*, *men*⟩ are extracted from 1149 and 38 alignments found in a subset of Web documents, out of which only four alignments are shown in the table. For example, the alignment of the sentence fragments *"to visit the city of their birth"* and *"to visit the place of their birth"* results in ⟨*city*, *place*⟩ becoming a potential paraphrase pair. On the positive side, the alignments capture properties shared among the potential paraphrases, such as the fact that both *cities* and *places* can be visited, located on a river, be lived in, or be the richest among others. Similarly, both categories of *dogs* and *men* can take steps, not be allowed to live somewhere, have an average age, and be killed or wounded. Unfortunately, the sharing of a few properties is not a sufficient condition for two concepts to be good paraphrases of each other. Indeed, neither ⟨*city*, *place*⟩ not ⟨*dogs*, *men*⟩ constitute adequate paraphrase pairs.

Arbitrary boundaries are oblivious to syntactic structure, and will often span only partially over otherwise cohesive linguistic units, such as complex noun phrases, clauses, etc. Their main limitation, however, is the lack of an anchoring context, that would act as a pivot to which the information within the sentence fragments would be in strong dependency. We argue that it is both necessary and

Table 2. Types of text anchors for sentence fragment alignment

Anchor Type	Examples
Named entities for appositives	(*"Scott McNealy, CEO of Sun Microsystems"*, *"Scott McNealy, chief executive officer of Sun Microsystems"*)
Common statements for main verbs	(*"President Lincoln was killed by John Wilkes Booth"*, *"President Lincoln was assassinated by John Wilkes Booth"*)
Common dates for temporal clauses	(*"1989, when Soviet troops withdrew from Afghanistan"*, *"1989, when Soviet troops pulled out of Afghanistan"*)
Common entities for adverbial relative clauses	(*"Global and National Commerce Act, which took effect in October 2000"*, *"Global and National Commerce Act, which came into force in October 2000"*)

possible to automatically extract anchoring context from the sentences, and use it in conjunction with the sentence fragments, to decide whether the fragments are worth aligning or not. Text anchors provide additional linguistic context to the alignment phase. Generally speaking, they are linguistic units to which the sentence fragments as a whole are in a strong syntactic or semantic relation. From the types of anchors suggested in Table 2, only the temporal relative clauses and the more general adverbial relative clauses are implemented in the experiments reported in this paper.

To ensure robustness on Web document sentences, simple heuristics rather than complex tools are used to approximate the text anchors and sentence fragment boundaries. Sentence fragments are either temporal relative clauses or other types of adverbial relative clauses. They are detected with a small set of lexico-syntactic patterns, which can be summarized as:

(Temporal-Anchors): *Date* [,|-|(|nil] *when TemporalClause* [,|-|)|.]
(Adverbial-Anchors): *NamedEntity* [,|-|(|nil] *WhAdv RelativeClause* [,|-|)|.]

The patterns are based mainly on *wh*-words and punctuation. The disjunctive notation [,|-|)|.] stands for a single occurrence of a comma, a dash, a parenthesis, or a dot. *WhAdv* is one of *who*, *which* or *where*, and a *NamedEntity* is approximated by proper nouns, as indicated by part-of-speech tags. The matching clause *TemporalClause* and *RelativeClause* 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.

The time complexity for brute-force pairwise alignment is the square of the cardinality of the set of sentence fragments sharing the same anchors. A faster implementation exploits an existing parallel programming model [16] to divide the acquisition and alignment phases into three extraction stages. Each stage is distributed for higher throughput.

Table 3. Examples of siblings within the resource of categorized named entities

Phrase	Top Siblings
BMW M5	S-Type R, Audi S6, Porsche, Dodge Viper, Chevrolet Camaro, Ferrari
Joshua Tree	Tahquitz, Yosemite, Death Valley, Sequoia, Grand Canyon, Everglades
NSA	CIA, FBI, INS, DIA, Navy, NASA, DEA, Secret Service, NIST, Army
Research	Arts, Books, Chat, Fitness, Education, Finance, Health, Teaching
Porto	Lisbon, Algarve, Coimbra, Sintra, Lisboa, Funchal, Estoril, Cascais

3.3 Categorized Named Entities for Paraphrase Validation

Spurious sentences, imperfect alignments, and misleading contextual similarity of two text fragments occasionally produce incorrect paraphrases. Another contribution of the paper is the use of a novel post-filtering mechanism, which validates the candidate paraphrase pairs against a large resource of InstanceOf relations separately acquired from unstructured Web documents.

The data-driven extraction technique introduced in [17] collects large sets of categorized named entities from the Web. A categorized named entity encodes an InstanceOf relation between a named entity (e.g. *Tangerine Dream*) and a lexicalized category (e.g., *progressive rock group*) to which the entity belongs. Both the named entity and the lexicalized category are extracted from some common sentence from the Web. Even though the algorithm in [17] was developed for Web search applications, it is exploited here as one of the many possible criteria for filtering out some of the incorrect paraphrase pairs.

The key source of information derived from the categorized named entities are the siblings, i.e. named entities that belong to the same category. They are directly available in large numbers within the categorized named entities, as named entities often belong to common categories as shown in Table 3. Since siblings belong to a common class, they automatically share common properties. This results in many surrounding sentence fragments that look very similar to one another. Consequently, siblings produce a significant percentage of the incorrect paraphrase pairs. However, these errors can be detected if the phrases within a potential paraphrase pair are matched against the siblings from the categorized names. If the elements in the pair are actually found to be siblings of each other, their value as paraphrases is questionable at best, and hence the pair is discarded.

4 Evaluation

4.1 Experimental Setting

The input data is a collection of approximately one billion Web documents from a 2003 Web repository snapshot of the Google search engine. All documents are in English. The sentence fragments that are aligned to each other for paraphrase acquisition are based on two types of text anchors. In the first run,

Table 4. Top ranked paraphrases in decreasing order of their frequency of occurrence (top to bottom, then left to right)

With Temporal-Anchors		With Adverbial-Anchors	
passed, enacted	percent, per- cent	died, passed away	included, includes
percent, per cent	took, came into	percent, per cent	played, plays
figures, data	totalled, totaled	United States, US	lives, resides
passed, approved	took, came to	finished with, scored	operates, owns
statistics, figures	over, more than	over, more than	consists of, includes
statistics, data	enacted, adopted	began, started	center, centre
United States, US	information is, data are	include, includes	came, entered
figures are, data is	information is, figures are	operates, runs	takes, took
statistics are, data is	was elected, became	begins, starts	lost, won
passed, adopted	statistics are, information is	effect, force	chairs, heads

Temporal-Anchors, the sentence fragments are relative clauses that are temporally anchored through a *when* adverb to a date. In the Adverbial-Anchors sun, the sentence fragments are adverbial relative clauses anchored to named entities through other *wh*-adverbs.

For each unique date anchor (Temporal-Anchors) and named entity anchor (Adverbial-Anchors), a maximum of 100,000 associated sentence fragments are considered for pairwise alignment to one another. The extracted paraphrase pairs are combined across alignments, and ranked according to the number of unique alignments from which they are derived. Pairs that occur less than three times are discarded.

The impact of the extracted paraphrases is measured on a test set of temporal queries. The set consists of 199 *When* or *What year* queries from the TREC Question Answering track (1999 through 2002) [18]. The queries extract direct results from an existing experimental repository of 8 million factual fragments associated with dates [17]. The fragments are similar to the sentence fragments from the Temporal-Anchors run, e.g., *1953* associated to *"the first Corvette was introduced"*, and *1906* associated to *"Mount Vesuvius erupted"*. Each query receives a score equal to the reciprocal rank of the first returned result that is correct, or 0 if there is no such result [18]. Individual scores are aggregated over the entire query set.

4.2 Results

Table 4 shows the top paraphrases extracted in the two runs, after removal of pairs that contain either only stop words, or any number of non-alphabetic characters, or strings that differ only in the use of upper versus lower case. A small number of extractions occur in both sets, e.g., ⟨*over*, *more than*⟩. At

Table 5. Quality of the acquired paraphrases computed over the top, middle and bottom 100 pairs

	Temporal-Anchors			Adverbial-Anchors		
Classification of Pairs	Top	Mid	Low	Top	Mid	Low
(1) Correct; synonyms	53	37	3	33	23	6
(2) Correct; equal if case-insensitive	4	7	0	9	2	14
(3) Correct; morphological variation	0	0	0	20	15	6
(4) Correct; punctuation, symbols,spelling	22	1	10	18	11	15
(5) Correct; hyphenation	2	33	0	2	19	43
(6) Correct; both are stop words	15	0	0	1	0	0
Total correct	96	78	13	83	70	84
(7) Siblings rather than synonyms	0	10	82	5	7	7
(8) One side adds an elaboration	0	11	4	4	3	1
Total siblings	0	21	86	9	10	8
(10) Incorrect; e.g., antonyms	4	1	1	8	20	8

least one of the pairs is spurious, namely ⟨*lost*, *won*⟩, which are antonyms rather than synonyms. Difficulties in distinguishing between synonyms, on one side, and siblings or co-ordinate terms (e.g., *Germany* and *France*) or even antonyms, on the other, have also been reported in [11]. The occurrence of the spurious antonym pair in Table 4 suggests that temporal anchors provide better alignment context than the more general adverbial anchors, as they trade off coverage for increased accuracy.

The automatic evaluation of the acquired paraphrases is challenging despite the availability of external lexical resources and dictionaries. For example, the lexical knowledge encoded in WordNet [7] does not include the pair ⟨*abduction*, *kidnapping*⟩ as synonyms, or the pair ⟨*start*, *end*⟩ as antonyms. Therefore these and many other pairs of acquired paraphrases cannot be automatically evaluated as correct (if synonyms) or incorrect (e.g., if antonyms) based only on information from the benchmark resource. To measure the quality of the paraphrases, the top, middle and bottom 100 paraphrase pairs from each run are categorized manually into the classes shown in Table 5. Note that previous work on paraphrase acquisition including [9], [13] and [16] also relies on manual rather than automatic evaluation components. The pairs in class (1) in Table 5 are the most useful; they include ⟨*photo*, *picture*⟩, ⟨*passed*, *approved*⟩, etc. The following categories correspond to other pairs classified as correct. For instance, ⟨*Resolution*, *resolution*⟩ is classified in class (2); ⟨*takes*, *took*⟩ is classified in class (3); ⟨*world*, *wolrd*⟩ is classified in (4); ⟨*per-cent*, *percent*⟩ in (5); and ⟨*has not*, *hasn't*⟩ in (6). The next three classes do not contain synonyms. The pairs in (7) are siblings rather than direct synonyms, including pairs of different numbers. 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 ⟨*complete data*, *records*⟩. Finally, the last class from Table 5 corresponds to incorrect extractions, e.g. due to antonyms like ⟨*started*, *ended*⟩. The results confirm that temporal anchors produce better paraphrases, at least over the first half of the ranked list of paraphrases. In comparison to the results shown in Table 5, the

Table 6. Examples of paraphrase pairs discarded by sibling-based validation

Discarded Pair	Ok?	Discarded Pair	Ok?
April, Feb.	Yes	Monday, Tuesday	Yes
season, year	Yes	country, nation	No
goods, services	Yes	north, south	Yes
Full, Twin	Yes	most, some	Yes
country, county	Yes	higher, lower	Yes
authority, power	No	Democrats, Republicans	Yes
England, Scotland	Yes	fall, spring	Yes

Table 7. Performance improvement on natural-language queries

Max. Nr. Disjunctions per Expanded Phrase	Nr. Queries with Better Scores	Nr. Queries with Lower Scores	Overall Score
1 (no paraphrases)	0	0	52.70
5 (4 paraphrases)	18	5	63.35

evaluation of a sample of 215 pairs results in an accuracy of 61.4% in [11], whereas 81.4% of a sample of 59 pairs are deemed as correct in [9].

The validation procedure, based on siblings from categorized names, identifies and discards 4.7% of the paraphrase pairs as siblings of one another. This is a very good ratio, if corroborated with the percentage of pairs classified as siblings in Table 5. Out of 200 pairs selected randomly among the discarded pairs, 28 are in fact useful synonyms, which corresponds to a projected precision of 86% for the validation procedure. Table 6 illustrates a few of the pairs discarded during validation.

The acquired paraphrases impact the accuracy of the dates retrieved from the repository of factual fragments associated with dates. All phrases from the test set of temporal queries are expanded into Boolean disjunctions with their top-ranked paraphrases. For simplicity, only individual words rather than phrases are expanded, with up to 4 paraphrases per word. For example, the inclusion of paraphrases into the query Q685: *"When did Amtrak begin operations?"* results in the expansion *"When did Amtrak (begin|start|began|continue| commence) (operations|operation|activities|business|operational)?"*. The top result retrieved for the expanded query is *1971*, which is correct according to the gold standard.

As shown in Table 7, paraphrases improve the accuracy of the returned dates, increase the number of queries for which a correct result is returned, and increase the overall score by 20%. Further experiments show that the incremental addition of more paraphrases, i.e., four versus three paraphrases per query word, results in more individual queries with a better score than for their non-expanded version, and higher overall scores for the returned dates. After reaching a peak score, the inclusion of additional paraphrases in each expansion actually degrades the overall results, as spurious paraphrases start redirecting the search towards irrelevant items.

5 Conclusion

Sophisticated methods developed to address various natural language processing tasks tend to make strong assumptions about the input data. In the case of paraphrase acquisition, many methods assume reliable sources of information, clean text, expensive tools such as syntactic parsers, and the availability of explicit document attributes. Comparatively, this paper makes no assumption of any kind about the source or structure of the input documents. The acquisition of paraphrases is a result of pairwise alignment of sentence fragments occurring within the unstructured text of Web documents. The inclusion of lightweight linguistic context into the alignment phase increases the quality of potential paraphrases, as does the filtering of candidate paraphrases based on a large set of categorized named entities also extracted from unstructured text. The experiments show that unreliable text of the Web can be distilled into paraphrase pairs of good quality, which are beneficial in returning direct results to natural-language queries.

Acknowledgments

The author would like to thank Péter Dienes for suggestions and assistance in evaluating the impact of the extracted paraphrases on natural-language queries.

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. Collins, M.: Head-Driven Statistical Models for Natural Language Parsing. PhD thesis, University of Pennsylvania, Philadelphia, Pennsylvania (1999)
5. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. In: Proceedings of the 38th Annual Meeting of the Association of Computational Linguistics (ACL-00), Hong Kong (2000) 512–520
6. 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
7. Miller, G.: WordNet: a lexical database. Communications of the ACM **38** (1995) 39–41

8. Turney, P.: Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In: Proceedings of the 12th European Conference on Machine Learning (ECML-01), Freiburg, Germany (2001) 491–502
9. 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
10. 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
11. 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
12. 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
13. 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
14. 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
15. 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
16. 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
17. Paşca, M.: Acquisition of categorized named entities for Web search. In: Proceedings of the 13th ACM Conference on Information and Knowledge Management (CIKM-04), Washington, D.C. (2004)
18. 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