Similarity Measures for Tracking Information Flow

Donald Metzler
University of Massachusetts
Amherst, MA 01003
metzler@cs.umass.edu

Yaniv Bernstein
RMIT University
Melbourne, Australia 3001
ybernste@cs.rmit.edu.au

W. Bruce Croft
University of Massachusetts
Amherst, MA 01003
croft@cs.umass.edu

Alistair Moffat
University of Melbourne
Melbourne, Australia 3010
alistair@cs.mu.oz.au

Justin Zobel
RMIT University
Melbourne, Australia 3001
jz@cs.rmit.edu.au

ABSTRACT

Text similarity spans a spectrum, with broad topical similarity near one extreme and document identity at the other. Intermediate levels of similarity – resulting from summarization, paraphrasing, copying, and stronger forms of topical relevance – are useful for applications such as information flow analysis and question-answering tasks. In this paper, we explore mechanisms for measuring such intermediate kinds of similarity, focusing on the task of identifying where a particular piece of information originated. We consider both sentence-to-sentence and document-to-document comparison, and have incorporated these algorithms into RECAP, a prototype information flow analysis tool. Our experimental results with RECAP indicate that new mechanisms such as those we propose are likely to be more appropriate than existing methods for identifying the intermediate forms of similarity.

Categories and Subject Descriptors:
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Experimentation, Theory

Keywords: Text reuse, information flow, statistical translation

1. INTRODUCTION

A text collection such as a newswire archive or web crawl typically contains a great deal of repeated information. Different authors may each present versions of a story or event; the same event may get presented in different ways for different audiences; and the facts of an event may get recapitulated each time it is discussed. Sometimes such presentations have little in common with each other but the broad subject matter; at other times one may be a copy of the other with minor edits.

Given a topic of interest, a sufficiently extensive archive may contain much of the history of the topic. In particular, the archive might plausibly be used to identify when particular ideas or statements first originated. In this work, our interest is in exploring whether we can identify alternative versions of the same information. It is not clear, however, that standard approaches to information retrieval or copy detection can be used for this task.

The extent to which passages of text are considered similar to each other can be thought of as lying somewhere on a *similarity spectrum*. At one end of this spectrum is identity; two documents that are the same as each other in every way clearly have the highest level of similarity possible. Discovery of such documents is the aim of systems for detecting plagiarism or co-derivation [Bernstein and Zobel, 2004; Broder et al., 1997; Heintze, 1996; Hoad and Zobel, 2003; Manber, 1994; Shivakumar and García-Molina, 1995]. The other end of the spectrum is the standard task of information retrieval: two documents are a match if they are topically related to the same information need.

Past research has largely focused on applications at one or other extreme of the spectrum, and there has been relatively little investigation of similarity tasks between these two extremes. However, in some applications, intermediate forms of similarity are of clear value. Examples include discovery of documents that summarize or paraphrase other documents, documents that are co-derived (that is, contain sufficiently similar material that they must at some point have come from the same source) and documents that share structure, or statements of fact.

In this paper, we explore the similarity spectrum in the context of our information flow analysis tool, RECAP [Metzler et al., 2005]. The objective of the RECAP project is to develop methods for tracking and analyzing the flow of facts and concepts through a text corpus. In order to create such a tool, we need a similarity measure that can reliably identify passages or sentences that share concepts and facts, that is, where information has been *reused*. This level of semantic resemblance is significantly stronger than simple topical similarity, but does not impose the syntactic similarity constraints typical of copy detection systems. Thus, we need to be able to accurately discern similarity in the middle portion of the similarity spectrum. Furthermore, the nature of the task requires that matches be evaluated at the sentence level, a further variation on the more usual document-to-document similarity measures.

We propose a range of approaches to reuse detection at the sentence level, and a range of approaches for combining sentence-level evidence into document-level evidence. To evaluate these approaches, we employ a hierarchy of five similarity levels: "unre-

CIKM'05, October 31–November 5, 2005, Bremen, Germany.
Copyright 2005 ACM 1-59593-140-6/05/0010 ...$5.00.

TREC DOCNO	Sentence
AP881019-0050	Its May 18, 1980, eruption leveled 230 square miles of evergreen forest, left 57 people dead or missing and sent up an ash cloud that circled the globe.
AP900107-0009	Mount St. Helens erupted May 18, 1980, leveling 230 square miles, leaving 57 people dead or missing and creating an ash cloud that circled the globe.
AP900112-0005	Mount St. Helens in Washington state erupted explosively in May 1980, levelling hundreds of square miles of forest, blowing ash so high into the atmosphere it circled the globe, killing 57 people and causing over $3 billion in damage.
AP900504-0193	*That was the day Mount St. Helens exploded, killing 57 people, flattening a forest and spitting out an ash cloud that circled the globe.*
AP900511-0075	The May 18, 1980, eruption of Mount St. Helens killed 57 people, devastated a vast area and lofted a huge ash cloud that circled the globe.
AP900511-0086	The May 18, 1980, eruption of Mount St. Helens killed 57 people, devastated a vast area and lofted a huge ash cloud that circled the globe.
AP901105-0146	Jonientz-Trisler said the explosion was "very minor, minor, minor" compared with a May 18, 1980, eruption that leveled 230 square miles of forest, left 57 people dead or missing and sent up an ash cloud that circled the globe.

Table 1: A list of sentences from TREC sources that demonstrate information reuse. The source sentence used for the query is italicized.

lated"; "on the general topic"; "on the specific topic"; "same facts"; and "copied". These levels can be regarded as points on the similarity spectrum.

Using relevance assessments at these similarity levels, we found marked differences in performance between the methods considered. The best methods were highly effective. Some of these methods were simple techniques that we introduced to establish whether such matching was feasible and whether it could be assessed. As a result, our research has a broad range of outcomes, including methods for reuse detection; measures of the quality of reuse detection; and a demonstration that reuse detection is both meaningful and feasible. As an aside, we have discovered that there is a high rate of reuse in the standard text collections used for information retrieval experiments.

Our work is to some extent exploratory rather than definitive, in that this problem has not been investigated before. On the other hand, our results show that even the preliminary methods we describe are suitable for reuse detection in practice.

2. TRACKING CONCEPT REUSE

Our research is motivated by the desire to develop effective methods for identifying and tracking idea and fact reuse within a collection. A user who is browsing a document should be able to select a sentence or group of sentences and be presented with a history of where the ideas and facts in the sentences were used elsewhere in the collection. In some cases, these ideas or facts may be alternative presentations of the same concepts; in other cases, much the same wording may be used, demonstrating that the statements have a common origin.

Such an application would be of great use to information analysts in both military and civilian fields. It would allow the origin and flow of specific facts and concepts through a text corpus to be analyzed and visualized. With additional information such as datestamps, such a tracking system could help establish when information was first known and in what context.

We have created a prototype tool for this task called RECAP. As an example of the uses of this software, suppose a user is browsing an article discussing the 1980 eruption of Mount St. Helens in Washington State. The user encounters the following sentence:

> That was the day Mount St. Helens exploded, killing 57 people, flattening a forest and spitting out an ash cloud that circled the globe.

The user is interested in finding other documents in the collection in which this information is used in the same way. Querying on this sentence with an appropriate tool should return the sentences shown in Table 1. These sentences illustrate different kinds of fact and concept reuse. Of possible interest to an information analyst is that the same idea appeared in an article that was published almost two years before the query article, using very similar wording. This may lead the analyst to the primary source for this fact or concept.

Note that all the sentences in Table 1 are clear examples of information reuse, and it is difficult to believe that these sentences were written independently of one another. However, there are significant differences in the presentation of information between the sentences. The degree of similarity between such passages is higher than simple topical overlap, but somewhat lower than the syntactic resemblance required by copy detection systems. This motivates us to explore the similarity spectrum, assessing the effectiveness of various techniques at detecting passage similarity at this level.

Furthermore, due to the nature of the application and also to the nature of the type of similarity we are hoping to find, we are initially interested in working at sentence-level, rather than document-level, granularity. Facts and ideas are, in general, cohesive structural units. If a fact is presented in a particular sentence in one document, we expect it to be presented similarly in a single sentence in a corresponding document. This is different to topical similarity, in which the semantic sense of topicality may be broadly spread across the entire document.

3. RELATED WORK

Several techniques have been devised for estimating similarity of text passages (typically whole documents) to each other.

Relative-frequency measures [Hoad and Zobel, 2003; Shivakumar and García-Molina, 1995] are a class of similarity functions based on comparison of relative frequency of word occurrences between two passages of text. Two identical passages have identical word frequencies relative to each other; insertions, deletions, and edits will slowly degrade the value of such a relative-frequency score. In both cases where such measures have been used, the aim has been to detect copying between whole documents.

Document fingerprinting [Brin et al., 1995; Broder et al., 1997; Heintze, 1996; Manber, 1994] techniques are also designed to detect copying. They operate by passing a fixed-length sliding window over a collection (a typical window size may be eight words) and storing a selection of these fixed-length *chunks* in an inverted index. Documents that have a number of such chunks in common are considered to be similar in the sense that there is copying.

The main difference between the various document fingerprinting techniques is in their choice of which chunks to index and which to discard. In general, unless no chunks are discarded, selection heuristics are lossy and systems are thus vulnerable to the possibility that all matching chunks between a pair of documents are discarded. The DECO system of Bernstein and Zobel used an efficient whole-collection analysis to discard only chunks that have no effect on determining similarity between documents [Bernstein and Zobel, 2004].

There are several standard approaches used for query-based information retrieval that may also be effective for reuse matching. Both vector-space and language modeling [Ponte and Croft, 1998] approaches are typically used for evaluating the relevance of a document to a given (usually short) query. Substituting a document for a query allows these methods to give an estimate of the similarity between two documents. Sanderson used a standard vector-space algorithm in exactly this way, using whole documents as queries in an attempt to find duplicate documents in a newswire collection [Sanderson, 1997].

Probably the main body of work concerning retrieval and comparison of text at the sentence level is that which addresses the TREC novelty track [Harman, 2002; Soboroff and Harman, 2003]. The TREC novelty track is a forum for promoting the identification of novel information in a result list. In the years that the track has run, the task has involved returning a list of sentences that are both relevant to a given query and novel with respect to the sentences that have come before it. As such, a successful attempt at this task must have an effective way of scoring sentences. Allan *et. al* analyze a number of different methods at the sentence level [Allan et al., 2003]. However, it is to be noted that the correspondence with our task is not exact: the portion of a novelty system that scores for relevance compares a sentence to a query, while the portion that compares sentences to sentences is attempting to score for novelty, which is a different – in fact, nearly opposite – notion to similarity.

The topic detection and tracking (TDT) initiative [Allan et al., 1998] is comprised of tasks in which similarity classification is required, sometimes at the sentence level. TDT consists of three classes of task: story segmentation, topic detection, and topic tracking. The first of these tasks, segmentation, require a stream of sentences from a news source to be divided up into distinct stories. Topic detection is an unsupervised learning task requiring stories discussing a new topic to be flagged as they come in, or the entire corpus to be retrospectively clustered by topic. Topic tracking is a supervised learning task in which stories must be assessed for membership of a number of predefined topics.

The previous work is of relevance to reuse detection from several perspectives. The segmentation task requires sentences to be assessed for similarity to other sentences in a story; topic detection and topic tracking both demand a level of similarity between documents in a cluster that is significantly stronger than broad topical overlap, and answers are expected to discuss the same event.

4. SENTENCE-LEVEL SIMILARITY

In this section we examine several approaches to evaluating the level of similarity between a pair of sentences. These are intended as a sample of the various methodologies that might be considered for evaluating text similarity.

All of the techniques calculate a similarity score $S(Q,R)$ between a *query sentence* Q and a *candidate sentence* R, intended to capture numerically the extent to which they convey the same information. The objective is to be able to calculate $S(Q,R)$ for all sentences R in a collection and know that when the score S is maximized, the sentence R has a high degree of similarity to the query sentence Q.

Word overlap measures

As a baseline measure we chose a simplistic word overlap fraction; that is, the proportion of words in Q that also appear in the candidate sentence R:

$$S(Q,R) = \frac{|Q \cap R|}{|Q|},$$

where $|Q \cap R|$ is the number of terms that appear both in Q and R. The intuition here is simple – if two sentences have many terms in common then they are likely to be similar to some degree.

We also experimented with a variant of the word overlap measure in which the score was adjusted to take inverse document frequency into account:

$$S(Q,R) = \frac{|Q \cap R|}{|Q|} \left(\sum_{w \in Q \cap R} \log \frac{N}{df_w} \right).$$

This models the fact that high IDF terms are typically stronger indicators of shared heritage between two sentences than are low IDF terms.

TF-IDF measures

TF-IDF measures are a broad class of functions used for estimating relevance and similarity typically between queries and documents. The fundamental intuitions are that the more frequently a word appears in a passage, the more indicative that word is of the topicality of that passage; and that the less frequently a term appears in a collection, the greater its power to discriminate between interesting and uninteresting passages.

Standard TF-IDF formulations, such as Okapi BM25 [Robertson et al., 1992], may not be appropriate here since we are focusing on sentence-level similarity. Therefore, we adopt the formulation used by Allan et al. for the TREC novelty track, which was shown to consistently – but not significantly – outperform language modeling based approaches for finding topically similar sentences [Allan et al., 2003]. The similarity function is:

$$S(Q,R) = \sum_{w \in Q \cap R} \log(tf_{w,Q} + 1) \log(tf_{w,R} + 1) \log \left(\frac{N+1}{df_w + 0.5} \right)$$

where $tf_{w,Q}$ is the number of times term w appears in query sentence Q; $tf_{w,R}$ is the number of times term w appears in a candidate sentence R; N is the total number of documents in the collection; and df_w is the number of documents that w appears in.

Relative-frequency measures

As discussed, relative-frequency measures have been shown to perform well at finding co-derivative documents. In this work we investigate how well such methods work at finding co-derivative pieces of text at the sentence level. We use a simple variation of the identity measure of Hoad and Zobel [2003]:

$$S(Q,R) = \frac{1}{1+\frac{\max(|Q|,|R|)}{\min(|Q|,|R|)}} \sum_{w \in Q \cap R} \frac{\log N/df_w}{1+|tf_{w,D} - tf_{w,R}|} ,$$

with the various quantities defined as above. The numerator is a standard IDF factor, while the denominator contains two parts, one designed to penalize inequalities in the relative frequency of a word between the two sentences, and the other to penalize differences in the overall lengths of the sentences.

Probabilistic models

Translation transforms text in one language to text in another, with the aim of preserving as much of the semantics of the original as possible. This is a reasonable model for the process that occurs when text is summarized, paraphrased, or otherwise has its facts and concepts reused, motivating the investigation of sentence similarity at the level of fact and concept reuse as an act of translation.

Statistical machine translation systems [Brown et al., 1993] aim to generate high-quality translations of sentences between natural languages. Such systems make use of parameterized statistical language models of both source and target language, and a parameterized *statistical translation model* that estimates the probability that a given target sentence is a translation of the source sentence. Given these models and a parameterization, the system searches a space of possible translations and returns the sentence with the highest probability.

We propose using statistical translation models in much the same manner to estimate the probability that one sentence is a translation of another. This translation probability will then serve as the basis of the similarity score for pairs of sentences.

Given an alignment A of corresponding words between the query sentence Q and target sentence R, and a distribution of term translation probabilities $P_t(\cdot \mid t)$, the probability of translation is calculated by taking a product of the translation probabilities of the aligned words:

$$P(Q, A \mid R) = P_l(|Q| \mid R) \prod_{i=1}^{|Q|} P_t(q_i \mid r_{a_i}) P(A \mid R) ,$$

where $P_l(|Q| \mid R)$ is the probability sentence R generates a translation of length $|Q|$; q_i is the i^{th} term in sentence Q; r_{a_i} is the term in sentence R that aligns to the i^{th} term in sentence Q; and $P(A \mid R)$ is the probability of an alignment given sentence R.

IBM's Translation Model 1 assumes that the alignment between words in the two sentences is equi-probable and that no length distribution is favored over any other. That is, both $P(A \mid R)$ and $P_l(|Q| \mid R)$ are uniform. After some algebraic manipulation, this leads to the following form for the similarity function:

$$S(Q,R) = \frac{1}{(|R|+1)^{|Q|}} \prod_{i=1}^{|Q|} \sum_{j=1}^{|R|+1} P_t(q_i \mid r_j) .$$

The original translation model assumes that each sentence has a special *null* term at position 1; this is the reason that the summation iterates through $|R|+1$ terms. The null term is used to represent the fact that the current term in Q does not align to any terms in R.

We make the distributional assumption that $P_t(q_i \mid r_1) = P(q_i \mid C)$, where C is the background model inferred from the collection as a whole. This proceeds from the intuition that – in the absence of any other evidence – an unaligned word is likely to be present in a sentence with a probability equal to its overall probability in the more generalized background language model. The probability of aligning to the null term dictates the influence of the background language model on the resulting translation. The uniform distributional assumption on alignment means that the effective probability that a term in Q will align to the null term is $1/(|R|+1)$. Here, we generalize the original model by assuming there exist μ null terms in each sentence, where μ is a non-negative integer. This results in each sentence having length $|R| + \mu$, where $|R|$ is the number of non-null terms in R. This model can be described as:

$$S(Q,R) = \frac{1}{(|R|+\mu)^{|Q|}} \prod_{i=1}^{|Q|} \left[\sum_{j=1}^{\mu} P(q_i \mid C) + \sum_{j=\mu+1}^{|R|+\mu} P_t(q_i \mid r_j) \right]$$

We now make the further simplifying assumption that each word translates to itself; that is, $P_t(q_i \mid r_j) = 1$ iff $q_i = r_j$. It can easily be shown that this results in the following form:

$$S(Q,R) = \prod_{i=1}^{|Q|} \frac{tf_{q_i,R} + \mu P(q_i \mid C)}{|R| + \mu}$$

giving precisely the language modeling query likelihood ranking function using Dirichlet smoothing with smoothing parameter μ [Zhai and Lafferty, 2001]. With $\mu = 1$, we get Berger and Lafferty's Translation Model 0 [Berger and Lafferty, 1999].

The models discussed here rely on strong simplifying assumptions, in particular, the assumption that every term only translates to itself. Given a good thesaurus, it may be possible to improve on these models by incorporating a more refined estimate of the true translation probabilities. At present, we have established and motivated a generalized and extensible framework for representing sentence-level similarity in terms of translation probabilities, and shown that the query likelihood model under various types of smoothing is precisely an instantiation of this framework. Other papers have shown such a connection, but have artificially introduced smoothing into the mix [Murdock and Croft, 2004]. We show that smoothing falls naturally out of the translation model itself under plausible assumptions. This provides a solid theoretical motivation for using this model for evaluating strong sentence-level similarity.

In this paper we explore $\mu = 1$ (Translation Model 0) and $\mu = 2500$ (query likelihood). The parameter μ can be viewed as a knob that allows us to control the type of queries (translations) given a high probability for some document. As μ approaches 0, the model becomes a coordinate level (word overlap) measure that will likely be good at finding exact matches. At the other extreme, as μ gets large more background terms are allowed, which is likely (and known to be) good at finding topically relevant matches.

5. SENTENCE-LEVEL EXPERIMENTS

In order to explore the similarity spectrum at the sentence level, we devised the six-point qualitative similarity rating scale shown in Table 2. We believe that this scale accurately covers the similarity spectrum, and allows us to experiment with different techniques and evaluate their effectiveness at different levels of similarity.

Category	Description
5	*The two sentences are identical modulo formatting.* American inventor Philo T. Farnsworth, a pioneer of television, was accorded what many believe was long overdue glory Wednesday when a 7-foot bronze likeness of the electronics genius was dedicated in the U.S. Capitol.
4	*All of the facts contained in the query sentence are conveyed in the candidate sentence and the candidate sentence is a minor revision of the query.* American inventor Philo T. Farnsworth, a pioneer of television, was honored when a 7-foot bronze likeness of the electronics genius was dedicated in the U.S. Capitol.
3	*All of the facts contained in the query sentence are conveyed in the candidate sentence and the candidate sentence is a non-trivial revision of the query.* With his 81-year-old widow, Elma Farnsworth, looking on, the inventor was extolled as the father of television and his statue was placed in the pantheon of famous Americans of the Capitol's National Statuary Hall.
2	*Some specific facts contained in the query sentence are conveyed in the candidate sentence.* The clear favorite was one Philo T. Farnsworth, an inventor who is considered the father of television.
1	*The two sentences are on same general topic.* If Utahans have their way, Philo T. Farnsworth will become a household name.
0	*The two sentences are unrelated.* The crew worked for more than two hours to separate the 8.5-foot bronze likeness of the city's fictitious boxer from the steps of the Philadelphia Museum of Art, which has repeatedly insisted it doesn't want the statue.

Table 2: Sentence-level judgment categories along with an example of a sentence in that category with reference to the query "*American inventor Philo T. Farnsworth, a pioneer of television, was accorded what many believe was long overdue glory Wednesday when a 7-foot bronze likeness of the electronics genius was dedicated in the U.S. Capitol.*"

In general similarity matching, applications would draw a similarity threshold at the boundary between categories 0 and 1, while copy-detection applications would define the boundary between categories 4 and 5. It is categories 2, 3 and 4, in which there is significant semantic, factual and structural overlap, that define the middle-ground of the similarity spectrum.

In order to evaluate the effectiveness of the various systems described above, we created a set of 50 single-sentence queries from the topics used by the TREC question answering track. Using RECAP, which is built on the Indri search engine [Metzler et al., 2004], the techniques described above were run against the full TREC newswire collection, which is composed of: *Associated Press* articles (1988-1990), the *Financial Times* (1991-1994), the *Los Angeles Times* (1989-1990), the *San Jose Mercury News* (1991), and the *Wall Street Journal* (1987-1992). The combined collection consists of 848,481 documents. All documents were stopped using a list of 418 common stopwords, but not stemmed.

The top 25 ranked results per query from each of the systems were placed into an evaluation pool, and each of the sentences independently judged by two of the authors. In cases where the judgments disagreed, the conflict was resolved by discussion. In total 2,711 individual sentences were judged; the breakdown of judgments by similarity category is presented in Table 3.

The judged sentences were then assigned to one of two categories – similar or non-similar. The threshold at which this assignment took place – that is, the similarity level required for a sentence to be considered similar – was varied for the experiments. The final "cumulative" column in Table 3 shows the number of sentences in the judged set that were considered relevant at each threshold. All runs on the 50 judged queries were then analyzed using the `trec_eval` tool.

Figure 1 shows the mean average precision (MAP) for the various techniques across the similarity thresholds. The graph shows that, for all the scoring methods, MAP increased as the similarity threshold became stricter. This is to be expected, as closely similar sentence pairs have more features in common that can be exploited by scoring techniques.

Category	Description	Count	Cumulative
5	Exact match	61	61
4	Minor revision	69	130
3	Major revision	193	323
2	Specific topic	762	1085
1	General topic	886	1971
0	Unrelated	740	
Total		2711	

Table 3: Distribution of sentence-level judgments.

Of more interest is the relative performance of the various scoring functions at a given similarity threshold. At the general topic level (similarities of 1 and above), query likelihood was clearly the best performer. This is not unexpected, as much past research has shown query likelihood to be effective at identifying topicality. At the specific topic level (levels 2 and over), query likelihood still had the highest MAP, although its relative advantage had lessened somewhat. At levels 3, 4 and 5, the relative performance difference between techniques was far smaller, but Translation Model 0 was consistently the most effective.

The TF-IDF and identity measures were consistently poor. The other four measures – word overlap, IDF-weighted word overlap, Translation Model 0, and query likelihood – were each at or near the highest level of effectiveness at one or more of the threshold levels tested. However, the relative effectiveness of these four scoring functions between levels 2 and 3 was reversed. This means that it is difficult to conclude that any one of the techniques is more effective in the middle similarity region. This might (or might not) be because all the functions tested were too closely modeled on

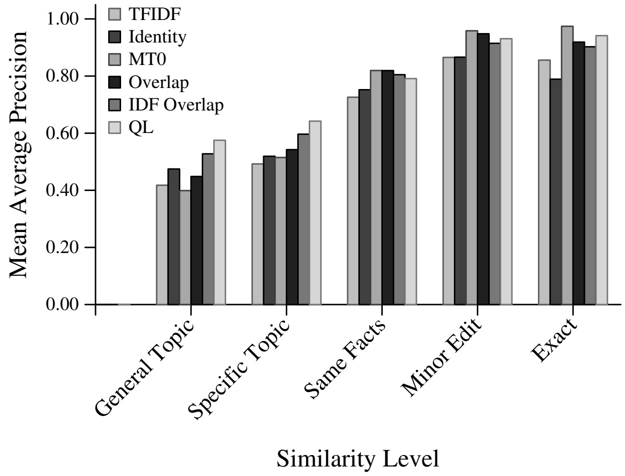

Figure 1: Mean average precision across similarity levels for sentence-level similarity measures.

techniques used for topical or syntactic similarity applications.

It is also worth noting that the baseline word overlap function was quite competitive on level 2 similarity and equal best for level 3. This further suggests that none of the techniques are performing at a particularly sophisticated level in this region of the similarity spectrum.

Nonetheless, the functions tested are sufficiently effective to render the RECAP tool useful, and, despite our suspicion that the none of the scoring functions was ideal, in absolute terms the MAP values are high enough to be used in a practical system. Because there is no clearly superior function, the current version of the RECAP software allows the user to choose which sentence-level scoring function to use.

6. DOCUMENT-LEVEL SIMILARITY

A key hypothesis in this investigation is that two documents that contain a significant overlap of facts and concepts can be expected to contain pairs of corresponding sentences that score highly at the sentence-to-sentence level. This assumption suggests building document scores by combining individual sentence-to-sentence scores in a bottom-up manner. The intuition is that, by examining semantic cohesion at the sentence level, we will be better able to distinguish document pairs that share a common body of facts and concepts from those that simply have a strong general topicality.

Another benefit of using a bottom-up approach is that the contribution of various sentences to the score is known, allowing corresponding blocks of concepts to be highlighted for the user. The RECAP system provides a slider that allows the user to set a sentence-level threshold to filter out matches that are insufficiently close.

We explored two different combination functions for calculating a bottom-up document similarity score $S(Q, D)$ between two documents Q and D. The first of these, SUM, is an exhaustive cross-alignment between all sentences in the two documents, similar in concept to Translation Model 1:

$$S(Q, D) = \prod_{q \in Q} \sum_{d \in D} S(q, d) P(d \mid D)$$

where q and d are sentences in the query and document, respectively, $S(q, d)$ is the similarity (or probability) between the query sentence s and document sentence d, and $P(d \mid D)$ is the likelihood (or weighting) of sentence d in D. In this case, all possible sentence scores will contribute to the final document similarity score. This means that, if a sentence has good correspondence to more than one sentence in the other document, all of these correspondences are able to make a contribution to the score. The disadvantage is that the many low scores caused by totally mismatching sentences will also contribute, possibly causing the function to be more susceptible to random noise.

Alternatively, we can base the document score on the best sentence matches:

$$S(Q, D) = \prod_{q \in Q} \max_{d \in D} S(q, d) P(d \mid D)$$

We call this combination function MAX. We note that it is possible that a sentence d in document D may be the best scoring match for two (or possibly more) different query sentences.

This removes both the advantage and the disadvantage discussed above as only the best possible match is counted towards the score. This scoring function can be thought of as taking the score once the sentences have been optimally aligned between the two documents.

Our early experiments showed that the second maximizing combination function consistently outperforms the first summing one, and it was adopted as the combination function for the experiments described in the next section. We also assumed that $P(d \mid D)$, the sentence-weight distribution, is uniform, although variations are possible.

7. DOCUMENT-LEVEL EXPERIMENTS

The aim of our next series of experiments was to examine the effectiveness of bottom-up document similarity scoring functions at evaluating document-level similarity in the middle of the similarity spectrum, in particular in identifying documents that share factual content. In the experiments we examine different sentence-level similarity measures $S(q, d)$. We compare these bottom-up schemes to a number of standard document-level similarity measures.

The methodology for these experiments was similar to that used for the experiment of sentence-level similarity functions, with the similarity spectrum for documents divided into the similarity levels described in Table 4.

For these experiments we used a set of 40 documents from the TREC newswire collection that were known to share facts and concepts with at least one other document in the collection. As with the sentence-level experiments, all techniques were run over the newswire collection for these 40 queries and the top 10 results aggregated into a pool for human judgment. This resulted in a pool of 1,538 documents to be judged, with the work shared between two judges. Table 5 summarizes the judgments that were made, and defines three similarity levels in terms of the categories listed in Table 4.

We chose two representative document-level techniques to evaluate as a basis for comparison. These techniques treat the entire document as a query rather than splitting it up into sentences. As previously discussed, the language-model derived query likelihood function [Ponte and Croft, 1998] has frequently shown to be effective at identifying topical similarity. It was also shown in Section 5 to be effective for sentence-level similarity, even at similarity thresholds in the middle of the similarity spectrum. The DECO system [Bernstein and Zobel, 2004] has similarly been shown to be effective and robust in detecting cases of syntactic similarity (text reuse). Thus, we have chosen as our baselines two techniques that have in the past been used for similarity assessment at the extreme ends of the similarity spectrum.

Category	Description
3	The two documents are identical, possibly except for minor edits; neither is a complete subset of the other.
2	There is sufficient overlap between (parts of) the two documents that there must have been common source material – for example, statement of identical numeric facts that would not be common knowledge, drawn from (for example) a press release.
1	There is some overlap between (parts of) the two documents, but not enough to conclude that the two authors had shared common source material – for example, because the shared content is "common knowledge".
0	There is no overlap between the two documents, and they are completely dissimilar.

Table 4: Document-level similarity judgment categories.

Category	Description	Count	Cumulative
3	Exact match	44	44
2	Reuse	212	256
1	General topic	275	631
0	Unrelated	907	
Total		2711	

Table 5: Distribution of document-level judgments.

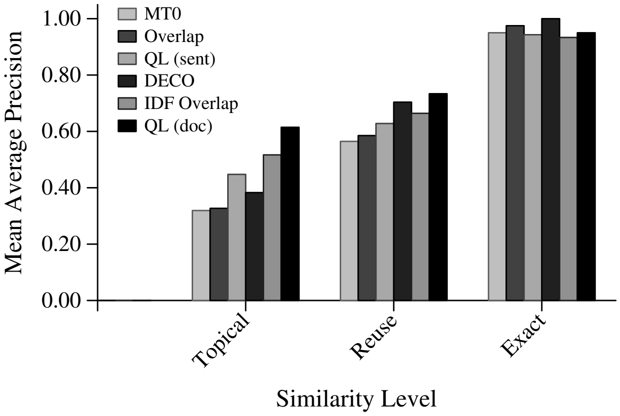

Figure 2: Mean average precision across varying types of similarity for each document-level measure.

In addition, we explored using the following sentence-level similarity measures in a bottom-up fashion: Translation Model 0 (mt0); unweighted word overlap (overlap); query likelihood (ql sent); and IDF-weighted overlap (idf overlap). The TF-IDF and relative frequency measures were not included because of their poor sentence-level effectiveness.

Figure 2 shows the MAP results for these experiments. Overall, the results for these experiments show that none of the bottom-up methods as tested were able to outperform both of the baseline all-of-document techniques at any of the three similarity thresholds.

Query likelihood at the all-of-document level was the most effective at levels 1 (topical similarity) and 2 (fact and concept reuse), while DECO was the most effective at level 3 (near identity) and second best at level 2. As expected, DECO was poor at detecting broader topical similarity.

Of the bottom-up measures, IDF-weighted overlap and query likelihood were the most effective for levels 1 and 2, whereas Translation Model 0 and unweighted word overlap were more effective when the threshold was set at level 3. Thus, more heavily smoothed measures did well when the similarity threshold was lower, whereas less smoothed measures were superior at matching documents that were near-identical. Not unexpectedly, for the bottom-up methods there was a strong correlation between scoring functions that performed well at the sentence level and the effectiveness of the measures based on these functions. This suggests that if better sentence-level scoring functions were to be devised, an immediate improvement in bottom-up scoring effectiveness would also result.

The inability of any of the bottom-up measures to significantly outperform the two standard all-of-document measures is disappointing. However, the difference in effectiveness – particularly when the similarity threshold is set at level 2 – is not large. There are two ways in which the bottom-up scoring methods can be made more effective. The first of these is to use a more effective scoring function at the sentence level. Improvements in the sentence level scores (as discussed above) will most likely flow on immediately to improved effectiveness at the document level. The second area in which improvements can be made is in the algorithms for aligning the sentences and combining the scores.

8. CONCLUSIONS

We have explored mechanisms for identifying passages of text that have varying degrees of similarity to a query passage. The ability to discern similarity between passages of text is valuable in a range of situations. Depending on the application, the threshold at which a pair of text passages are considered similar may vary. In general, a scoring technique that can effectively identify similar documents at one threshold of similarity might not be effective for a different similarity threshold, so the similarity threshold appropriate to an application plays an important role in determining an appropriate scoring technique.

Much past research has focused on finding effective techniques for quantifying similarity at one or other extreme of the similarity spectrum – either broad topical similarity, or strong syntactic correspondence. Little research has focused on the intermediate points of the similarity spectrum, between these two extremes. In this paper we examined some of the issues involved in more thoroughly exploring these types of similarities, via the use of RECAP, a prototype tool for analyzing fact and concept reuse.

Our experiments on different similarity measures at the sentence level led to discovery of several reasonably effective matching methods, in particular those that are based on a simplification of the probabilistic translation model paradigm. However, no one technique was able to significantly outperform the baseline measure, which was a simple measure of word overlap.

Our experiments with techniques that combine sentence-level scores in a bottom-up fashion for scoring document-to-document similarity showed that use of sentence-level evidence is a promising area of future work. A particular benefit of our bottom-up approach is that it allows easy localization and presentation of possible matches, and can be computed relatively efficiently. There are several avenues that may yield improved bottom-up methods. In particular, we intend to investigate aggregation and alignment methods from genomic search that may provide more precise scoring functions.

However, our experiments – and our experience with RECAP – have demonstrated that tracking of information reuse is practical and meaningful. The simple measures we have developed so far are able to accurately locate alternative instances of passages and, combined with timestamps, help a user to identify where a piece of information originated within a corpus. Further research may refine these methods, but they are already sufficently effective for use in practice.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by Advanced Research and Development Activity and NSF grant #CCF-0205575, in part by the Australian Research Council, and in part by the ARC Center for Perceptive and Intelligent Machines in Complex Environments. Any opinions, findings and conclusions or recommendations expressed in this material are the authors and do not necessarily reflect those of the sponsors.

References

J. Allan, A. Bolivar, and C. Wade. Retrieval and novelty detection at the sentence level. In *Proc. 26th Ann. International ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 314–321, 2003.

J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

A. Berger and J. Lafferty. Information retrieval as statistical translation. In *Proc. 22nd Ann. International ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 222–229, 1999.

Y. Bernstein and J. Zobel. A scalable system for identifying coderivative documents. In *Proc. String Processing and Information Retrieval Symp.*, pages 55–67, 2004. Published as LNCS 3246.

S. Brin, J. Davis, and H. García-Molina. Copy detection mechanisms for digital documents. In *Proc. ACM SIGMOD Ann. Conf.*, pages 398–409, 1995.

A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the web. *Computer Networks and ISDN Systems*, 29(8-13):1157–1166, 1997.

P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

D. Harman. Overview of the TREC 2002 novelty track. In *Proc. 11th Text REtrieval Conf. (TREC 2002)*. NIST, 2002.

N. Heintze. Scalable document fingerprinting. In *Proc. USENIX Workshop on Electronic Commerce*, November 1996.

T. Hoad and J. Zobel. Methods for identifying versioned and plagiarised documents. *Journal of the American Society of Information Science and Technology*, 54(3):203–215, 2003.

U. Manber. Finding similar files in a large file system. In *Proc. USENIX Winter Technical Conf.*, pages 1–10, San Fransisco, CA, USA, 17–21 1994.

D. Metzler, Y. Bernstein, W. B. Croft, A. Moffat, and J. Zobel. The RECAP system for identifying information flow. In *Proc. 28th Ann. International ACM SIGIR Conf. on Research and Development in Information Retrieval*, Aug. 2005. Demonstration abstract, to appear.

D. Metzler, T. Strohman, H. Turtle, and W. B. Croft. Indri at terabyte track 2004. In *Proc. 13th Text REtrieval Conf. (TREC 2004)*. NIST, 2004.

V. Murdock and W. B. Croft. Simple translation models for sentence retrieval in factoid question answering. In *Proc. SIGIR Workshop on Information Retrieval for Question Answering*, pages 31–35, 2004.

J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proc. 21st Ann. International ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 275–281, 1998.

S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at TREC. In *Proc. 1st Text REtrieval Conf. (TREC 2001)*, pages 21–30. NIST, 1992.

M. Sanderson. Duplicate detection in the Reuters collection. Technical Report TR-1997-5, University of Glasgow, 1997.

N. Shivakumar and H. García-Molina. SCAM: A copy detection mechanism for digital documents. In *Proc. 2nd Conf. on the Theory and Practice of Digital Libraries*, 1995.

I. Soboroff and D. Harman. Overview of the TREC 2003 novelty track. In *Proc. 12th Text REtrieval Conf. (TREC 2003)*, pages 38–53. NIST, 2003.

C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad-hoc information retrieval. In *Proc. 24th Ann. International ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 334–342, 2001.