Building a Resource for Evaluating the Importance of Sentences

Barry Schiffman

Columbia University
Department of Computer Science
New York, N.Y. 10027, U.S.A.
bschiff@cs.columbia.edu

Abstract

This paper will introduce a new lexical resource for measuring the importance of short segments of text, such as sentences. The resource, a list of words compiled automatically from a large background corpus of news articles, can provide evidence that a text segment is globally important, that is intrinsically interesting, not only interesting in relation to a specified topic or set of documents.

1. Introduction

A means of automatically identifying important passages would be valuable across a large number of natural language processing applications. The notion of importance in this paper is more general than what researchers often mean. It is importance in a general sense, without respect to a particular document or set of documents—a kind of global importance. An important passage would be one that people would notice, that they would recognize as depicting an interesting event. In contrast, the idea of importance that is often used in natural language processing is tied to some condition set out in the problem. Here are several examples: The document, or set of documents in automatic summarization, or the query in information retrieval and web searching, or the question in question-and-answer systems. In such cases, the problem is narrowed to a kind of matching exercise: words in the summary to the original documents; or, in question-answering or web-searching, terms to the corpus. But how does one assess importance without such yardsticks? How does one recognize that a passage would be interesting?

The quality I am working to capture is central to the discipline of journalism. It is the distinction between those events in the real world that make news and those that don't. This research assumes that professional journalists know what kinds of events and information their readers want and need to know. It also assumes that journalists intentionally try to convey the import and interesting aspects of their articles with the kind of language they use.

The difficulty is to decide how to analyze news reports in an effective way with the language processing tools available today. The experiments presented here are based on features that are easy to extract—the noninflected forms of content words that are found in the first paragraphs of news articles. So rather than attack the more difficult and subjective problem of what topics make the news, I set out to capture what language, specifically what words, are used to express the news that journalists usually emphasize by the placement at the beginning of their reports. The result is a dictionary of nouns, verbs and adjectives that are indicators of passages that readers would likely recognize as important or interesting—those with high *Lead Values*. In this paper, I will next discuss the two research efforts that lead to this inquiry, related work, the experiments I conducted and my efforts to evaluate the results.

2. Applications

I have been working on a multidocument summarization system which seeks to contrast the documents in a collection as well as compare them. One module in this system will identify *new information* and another will produce a summary from the information in the documents. Both tasks require the recognition of differences as well as commonalities, and I have found that typical approaches to term weighting falls short of the requirements. Both modules are intended to be integrated into the COLUMBIA UNIVERSITY NEWSBLASTER system (McKeown et al., 2002). The summarizer, DEMS or Dissimilarity Engine for Multidocument Summarization, is currently one of two summarizers in the system, and the one that processes the clusters of news articles that are more diverse and cover a larger span of time (Schiffman et al., 2002).

The need for an importance metric arose first in the exploration of the *new information detection* task, in which a system would continually monitor incoming reports on an event or larger topic, and try to discern what information is new, that is, what has not yet been seen. It quickly became apparent that most segments of the incoming documents contained some new information, and it was necessary to find a way for deciding:

1. Whether the new passage is substantially different in meaning from what has already been seen or whether it is merely written in a different way.

2. If there is a substantial difference in meaning, is it important enough to selected and shown to the user.

If these two tasks are not accomplished, the system will fail by giving the user nearly everything. This problem was encountered by a group of researchers at the Johns Hopkins Summer School who began exploring the *new information detection* task in 1999 (Allan et al., 1999). One of the participants, James Allan, later wrote that they found about 80% of the sentences in their sets of articles contained some new information Allan et al. (2001).

I encountered the same phenomenon in my own experiments in new-information detection experiments and began trying to devise some measure to capture the qualities

of important and interesting. The system is still under development and will eventually use the *Lead Value* metric as well as others to select passages for update summaries (Schiffman and McKeown, 2002).

The problem arose again in the multidocument summarization evaluation at the Document Understanding Conference (Harman and Over, 2002). The DUC document sets each consisted of about 12 news articles that were often only loosely tied together, spanning several years. General-purpose summaries were required without any clear way of deciding what to include.

The DEMS summarizer was developed to handle the DUC sets. The people who wrote the summaries against which the automatic summaries are evaluated were free to devise a theme for each article set, rewrite as much as they liked and organize the summary in any way. Like DEMS, most automatic summaries use sentence extraction because of the lack of robust technology to efficiently break down passages. To deal with the relatively chaotic problem, DEMS used a number of features to determine the qualities I outlined above, global importance and intrinsic interest, and the *Lead Value* metric was one of them.

3. Related Work

3.1. Machine Learning

A number of researchers in machine learning have used existing corpora that are in some way preselected or partially annotated for one purpose or another. In an information extraction experiment Mark Craven at CMU Craven (1999) used what he called "weakly" labeled data to reduce the cost of annotating a training corpus. He was seeking a way to map medical texts into a structured data base. He used a database that contained links to related text articles. Thus he could automatically collect relevant articles, reducing the effort to prepare his training corpus.

Another group working on information extraction at CMU, Seymour and others, sought to build a database of information about computer scientists from "distantly labeled data" composed of the header information on research papers. They reported high accuracy with Hidden Markov Models trained over this prefabricated data (Seymore et al., 1999).

Ellen Riloff learned textual-syntactic patterns for information extraction by comparing two corpora, a target containing the information she was interested in, and the other a general corpus (Riloff, 1996). The idea is that patterns of specialized words and syntactic structures will show up in greater numbers in the target corpus than in the general corpus.

My experiment is similar to all three in that I am considering the first paragraphs of news articles as a preselected corpus of *important* and *interesting* information. But my system differs from the two CMU groups, which use a more structured kind of data, and from the Riloff work, which is seeking to find very specific patterns from a very loose corpus.

3.2. New Information

The research in *New Information Detection* closest to mine is being done at the University of Massachusetts; their group seeks to produce a summary of related events as they change over time (Allan et al., 2001). They posit that a sentence is "useful" if it is on topic, and that a sentence is "novel" if it is not redundant with previously seen sentences.

Their perspective is topic-based and the experimental corpus comes from the TDT-2 corpus, in which 60,000 news stories were assigned to some 200 news topics. After selecting 22 of these topics, annotators created lists of the events that comprised each topic and assigned each sentence to one or another event. A total of 343 events were derived from 944 articles. Two different language models for deriving "useful" information were developed, based on the probabilities that individual words of a sentence appear in on-topic sentences or articles. The models of novelty are derived in a similar way from the specific words in on-event sentences.

Their notion of "useful" stands in sharp contrast to what I mean by "globally" important or "intrinscally" interesting. Their two measures, which are based on essentially the same primitives, risk canceling each other out. I am proposing to test importance independently with *Lead Value* and other metrics.

3.3. Summarization

The closest summarization system in spirit to DEMS is the NEATS multidocument summarization system (Lin and Hovy, 2001) which uses topic signatures, which try to discern the most frequently occurring topics in a document. While DEMS uses similar metrics, it adds the *Lead Value* feature to try to locate passages that might add something new and different to the summary, provided it is interesting enough.

A number of systems measure similarity between sentences and give greater weight to those that are most repetitive, making the assumption that repetition is an indicator of importance. Systems that do so include Multigen (McKeown et al., 1999), which was also developed at Columbia University, the University of Texas system (Harabagiu et al., 2001), focusing on information extraction techniques, and the ISI system (Marcu, 2001), which used discourse structure. A group at CMU (Goldstein et al., 2000) uses cosine similarity of vectors in the MMR algorithm. A graph representation of several relationships between words is used to find similarities and differences between pairs of articles (Mani and Bloedorn, 1997).

DEMS emphasizes statements that are different by treating importance as a separate issue. Although it computes a metric on how often different concepts, which are defined as sets of semantic equivalents, appear, there is no weight to similar passages.

4. Experiment

4.1. Journalism

It is an established practice in journalism to devote a large amount of attention to the "leads" of articles. It is based on the realization that news consumers spend only a limited time reading the news. The writers and editors of the news staunchly believe that they must win over the readers in the headline and first paragraph. Since most articles

cynical	coaxing	eerie	renovator	cling	impressionism	cutter
impressionist	conscription	tusker	ammonia	worn-out	convalescent	vial
unplayable	waterlogged	syphilis	decathalon	dragonfly	gigantic	showpiece
extricate	unbowed	cherry	waterborne	watershed	phenobarbital	reappearance
rivet	westernmost	heady	beloved	placid	bloke	caravan
large-scale	windfall	petrol	dame	mend	truffle	gutsy
chubby	enthral	enunciate	dank	chunk	stopgap	freak
intrusion	pensive	meld	mortuary	well-kept	well-established	one-man
linguist	zealotry	impresario	ostrich	possess	chump	crestfallen
menu	electronics	nationalize	restive	daub	crowning	vile
wizard	finalist	dishevelled	crossroad	autism	East	workable
reverberate	excitable	trawler	sizeable	insolvent	stewardess	rhyme
fluorescent	sharpen	spectre	infighting	setter	electrical	mesa
jeopardize	rude	rambunctious	polyglot	chivalry	statistical	bloodbath

Table 1: A sampling of the *Lead Values* from the 1996 Reuters news wire. They are kept in the order they were placed in a hash table; they are also used only as binary values.

are straightforward accounts of factual events, journalists like to make it clear at once why their subjects are interesting and important enough to read.

Rau, et al., in a 1994 paper on summarization found that the first paragraph of a news article often served as an excellent summary. Of course, the summarization problem is more complicated. For one, an article may address more than a single topic, or a longer summary might be desired. In multidocument summarization, the first-paragraph technique may well produce redundant summaries, or summaries of the wrong length. Finally, a substantial number of articles are "human interest" articles where the writers use more artistic language in their "leads," hoping to draw in readers and the explanation of why the article is interesting and important is delayed for several paragraphs.

4.2. Method

The method in this work tries to discover features about the first paragraphs, or the "leads", that could be used to identify important information anywhere in the document. My method is partly inspired by the researchers in information extraction who try to make use of partially marked data to build training corpora for machine learning. This experiment takes the simplest approach by considering just the words.

In this case the premarked data are contained in a large corpus of news. The uninflected forms of each content word is looked up, and counted. The features of lead paragraphs and the features of entire articles – in both cases just words – are compared, searching for those that could identify *lead paragraphs*, which I am considering a stand-in for *important passages*. The likelihoods of content words appearing in the leads and in the entire articles were collected, and the ratios were examined. A *Lead Value* is defined as one that tends to occur more often in the "leads" than in the article as a whole:

$$\frac{p(W_{inlead})}{p(W_{anywhere})} > 1$$

The corpus used was 38 million words of Reuters news wire from 1996, and a lexicon of 4,997 *Lead Values* was derived. Thus the corpus is interpreted as a partially annotated collection of articles. Table 1 shows a list of 98 of the words, most of which are words of some impact, like "intrusion", "extricate", "zealotry", and "watershed". Others are a bit puzzling, but are likely to be related to events that were much in the news in 1996. I intend to collect similar samples, from other years and other sources, and to refine the list by using those words that are consistently found in the "leads."

The ratios were checked for statistical significance with the binomial test and only those with ratios where $pvalue < 0.05$ were accepted for inclusion in the dictionary.

Since the collection of the raw data is both simple and reliable, few errors are being put into the lexicon because of extraction failures.

5. Evaluation

5.1. Task-Based

The lexicon is being used successfully in a multidocument summarization system, DEMS, which is used daily by COLUMBIA UNIVERSITY NEWSBLASTER, and was evaluated at the Document Understanding Conference in 2001. The conference did not set a single metric to compare systems, but it was clear that DEMS was among the top systems, although evaluation of such subjective tasks is always problematic (McKeown et al., 2001).

In DEMS, the *Lead Value* feature is one of 11 features in ranking sentences for inclusion in the summary. Six of the features are word based, and *Lead Value* is weighted the strongest among the six. The other five features are some that are specific to journalism, like the location of the sentence and the publication date, and others that pertain to linguistic features, like penalties for sentences that are either too short or too long.

To illustrate *Lead Value* and compare it to other measures, I will use a single news story and show which paragraphs had the highest average values for *Lead Value*, *Word*

1. DETROIT (Reuter) - In what defense attorneys are billing "the trial of the century," right-to-die advocate Jack Kevorkian returns to the courtroom Tuesday to face charges of violating Michigan's controversial suicide law.
2. Kevorkian, 65, is certainly no stranger to the courts. During the last three years, the retired pathologist has appeared in front of numerous judges to defend the right of terminally ill people to end their lives under his care.
3. But Tuesday Kevorkian gets a chance to tell his story to a jury for the first time.
4. "I think this is the trial of the century," said Kevorkian attorney Geoffrey Fieger. "I think this will become the world's most famous court trial."
5. Since Kevorkian first started helping sick people commit suicide in 1990, public opinion on the issue has been sharply divided in Michigan and the rest of the country.
6. So far, Kevorkian has helped 20 people end their lives.
7. Last year, the Michigan legislature, in an attempt to stop Kevorkian from participating in any more suicides, passed a law making it a felony, punishable by up to four years in prison and a $2,000 fine.
8. Since then judges in three cases have ruled the law is unconstitutional and have thrown out charges against Kevorkian. But a fourth and final charge still remains.
9. Unlike many of Kevorkian's previous cases, in which he declined to cooperate with authorities or testify in court, the retired pathologist has openly admitted to helping Thomas Hyde commit suicide by inhaling carbon monoxide gas.
10. Hyde, 30, suffered from amytrophic lateral sclerosis (ALS), better known as Lou Gehrig's disease. He died August 4 in the back of Kevorkian's van on Belle Isle, near Detroit.
11. Fieger said Kevorkian made the admission to force a showdown in the courts.
12. "We're putting the reactionary forces in society that got this law passed in the first place on trial," Fieger said. "This is a trial about the right of individuals not to suffer, period. And we will win."
13. On Monday Fieger will ask Detroit Recorders Court Judge Thomas Jackson to allow him to enter into evidence a gut-wrenching videotape of Hyde, made a month before his death.
14. Hyde, who went from being physically fit to a near invalid in a matter of months, could barely speak during the session. But near the end of the videotape he finally utters "I want to end this. I want to ... die."
15. Wayne County Prosecutor Timothy Kenny has filed a motion to oppose the move on the grounds that the only issue to be decided by the jury is whether Kevorkian broke the law.
16. "This trial is about obeying the law," Kenny said. "The law is the law."
17. The Michigan Court of Appeals is currently reviewing the controversial suicide law on constitutional grounds. Prosecutors were willing to wait for the higher court to issue its ruling, but Fieger said Kevorkian wants the issue to be decided by a jury.
18. "We need to have a trial," Fieger said. "The people need to know that this isn't simply a matter in which the courts are going to usurp, absolutely, the power of the people.
19. "There are going to be 12 jurors up there who are reflective of the conscience of society, and they are going to acquit Dr Kevorkian."
20. Fieger said he plans to allow Kevorkian to testify.
21. The only other assisted suicide case in Michigan to go before a jury ended in an acquittal in 1991, when a Detroit Recorders Court Jury failed to convict Bertram Harper for helping his wife commit suicide by putting a plastic bag over her head.
22. Harper, 72, was charged with murder because the state did not have a suicide law on the books then.
23. Kenny was the prosecutor in that case as well.
24. Pretrial motions will be heard Monday, with jury selection scheduled to begin Tuesday.
25. The trial is expected to last about a week.

Figure 1: A Reuters article about Dr. Jack Kevorkian in 1994 that was written as a preview of a prominent court battle. It is not a "hard news" story and the first paragraph is very general.

Frequency[1] and *TF/IDF*[2] The first two are heavily weighted in DEMS. The last one is not used by the summarizer, but I included it here because it is so widely used in a large number of tasks. (Others have found that it has not been helpful in multidocument summarization (Lin and Hovy, 2000).)

Showing complete summaries would not be as helpful since the characteristics measured by the features are not independent of one another, so it is impossible to isolate the value of one of them, especially in view of the fact that there is no established standard against which automatically produced summaries can be evaluated. Further, the system is also intended to be a multidocument system that must select only a few sentences from a large number. In the DUC evaluation, the typical set had 10 or 12 articles, yet the shortest summaries allowed for only two or three average-size sentences. The combination of *Lead Value* and sentence location tend to pick the punchiest of the actual leads.

[1]Counts of the words, but in *DEMS* equivalent words are grouped together and the counts are done over the document set.

[2]Term Frequency-Inverse Document Frequency, a metric that comes from Information Retrieval research on distinguishing documents related to a query.

By *Lead Values* 10. Hyde, 30, suffered from amytrophic lateral sclerosis (ALS), better known as Lou Gehrig's disease. He died August 4 in the back of Kevorkian's van on Belle Isle, near Detroit. 13. On Monday Fieger will ask Detroit Recorders Court Judge Thomas Jackson to allow him to enter into evidence a gut-wrenching videotape of Hyde, made a month before his death.
By *TF/IDF* 8. Since then judges in three cases have ruled the law is unconstitutional and have thrown out charges against Kevorkian. But a fourth and final charge still remains. 5. Since Kevorkian first started helping sick people commit suicide in 1990, public opinion on the issue has been sharply divided in Michigan and the rest of the country.
By *Word Frequency* 17. The Michigan Court of Appeals is currently reviewing the controversial suicide law on constitutional grounds. Prosecutors were willing to wait for the higher court to issue its ruling, but Fieger said Kevorkian wants the issue to be decided by a jury. 19. "There are going to be 12 jurors up there who are reflective of the conscience of society, and they are going to acquit Dr Kevorkian."

Table 2: The pairs of paragraphs that had the best average value for the three metrics, *Lead Value*, *TF/IDF* and *Word Frequency*

The example article is about Jack Kevorkian, the doctor who assisted people in committing suicide (See Figure 1), the "lead paragraph" says a new court case will begin, but the specific point of the article is not clear until the 9th paragraph, which begins the segment that contains the top two passages as ranked by *Lead Value* (See Table 2). These two are highly specific and dramatic.

The passages selected by *TF/IDF* provide interesting perspective and background but they do not address this specific event. They give a different kind of perspective, and are related to this article, but not at the heart of what is happening.

5.2. Intrinsic

If important information in news articles does indeed tend to appear near the beginnings of articles, then an importance metric should be able to locate many of those paragraphs by giving them high ratings. I am again using partially marked data as a test corpus: A previously unseen collection of news articles – that is already partitioned into "important" segments and "other" segments – is used as standard.

One way to do this would be to compute the average location of the paragraphs ranked as the most important by the different metrics. In such a test (Table 3), the *Lead Value* feature performed about as well as *Word Frequency* and both were better predictors of "important" paragraphs than the TF/IDF metric. And a combination of *Lead Value* and *Word Frequency* did better than either one in isolation. Note that the *Lead Value* list was drawn from the Reuters collection for 1996, while the test used a random selection of 1,632 articles from the Reuters for 1994 to avoid possible bias from the events of 1996. By contrast, TF/IDF values were drawn from the 1994 corpus.

Importance Metric	Average Index
Random Baseline	6.186
TF/IDF	5.208
Document Frequency	4.545
Lead Values	4.596
Combination[†]	4.397

Table 3: The table shows the average paragraph location predicted by the measures I tested. The indices begin at 0. The optimal value is not known, but one would expect that it would tend to be in the beginning part of the article. [†]The weights for the combination were 0.09 for lead words and 0.91 for document frequency.

In a variation, I measured how often the actual "lead paragraph" was included in a summary, when paragraphs are chosen for the summary exclusively by the different importance metrics (Table 4).

6. Conclusion and Future Work

The contribution in this research is to present an innovative measure of importance that can be valuable in analyzing text automatically. I am not claiming that the *Lead Value* metric is a complete model of importance, but rather

Importance Metric	Percentage Included
TF/IDF	53.0
Word Frequency	67.2
Lead Values	78.9

Table 4: How often the "lead" paragraph is included in a summary compressed to 20% of its original size.

it can be used as evidence that a particular passage is important. That is how I use the metric in the DEMS summarizer, and in continuing research into new information detection. The *Lead Value* lexicon is currently based on one-year's worth of news reports from one news source, and it should be expanded. I have also examined the New York Times news reports, but paragraphs tended to be much longer and it is not clear how to normalize the units.

The usefulness of this narrowly focused lexicon suggests that other observations about text might be useful in discovering words that carry some impact and importance, for example, headline words, which could be readily collected. Beyond the news genre, the language enclosed in various html tags, like "Hn" or "B" or "U" might serve the same purpose.

In a larger sense, one might search for many narrowly defined characteristics of language in a large corpus, and produce useful resources as long as the means to accomplish accurate extraction exists.

7. References

James Allan, Hubert Jin, Martin Rajman, Charles Wayne, Daniel gildea, Victor Lavrenko, Rose Hoberman, and David Caputo. 1999. Topic-based novelty detection, 1999 summer workshop at clsp, final report. Technical report, Johns Hopkins University.

James Allan, Rahul Gupta, and Vikas Khandelwal. 2001. Temporal summaries of news topics. In *Proceedings of the ACM-SIGIR Conference*.

Mark Craven. 1999. Learning to extract relations from medline. In *AAAI-99 Workshop on Machine Learning for Information Extraction*.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of ANLP/NAACL-2000 Workshop on Automatic Summarization*.

S. Harabagiu, D. Moldovan, P. Morarescu, F. Lacatusu, R. Mihalcea, V. Rus, and R. Girju. 2001. Gistexter: A system for summarizing text documents. In *Proceedings of the Document Understanding Conference (DUC01)*.

Donna Harman and Paul Over. 2002. The duc summarization evaluations. In *Proceedings of the Human Language Technology Conference*.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th Annual International Conference on Computational Linguistics*.

C-Y Lin and E. Hovy. 2001. Neats: A multidocument summarizer. In *Proceedings of the Document Understanding Conference (DUC01)*.

Inderjeet Mani and Eric Bloedorn. 1997. Multi-document summarization by graph search and matching. In *Proceedings, American Association for Artificial Intelligence 1997*.

D. Marcu. 2001. Discourse-based summarization in duc-2001. In *Proceedings of the Document Understanding Conference (DUC01)*.

Kathleen McKeown, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards multidocument summarization by reformation: Progress and prospects. In *Proceedings of American Association for Artificial Intelligence 1999*.

K.R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, B. Schiffman, and S. Teufel. 2001. Columbia multi-document summarization: Approach and evaluation. In *Proceedings of the Document Understanding Conference (DUC01)*.

Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Carl Sable, Barry Schiffman, and Sergey Sigelman. 2002. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the Human Language Technology Conference*.

Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of the 13th National Conference on Artificial Intelligence*.

Barry Schiffman and Kathleen McKeown. 2002. Towards the identification of new information, submitted to naacl 2002.

Barry Schiffman, Ani Nenkova, and Kathleen McKeown. 2002. Experiments in multidocument summarization. In *Proceedings of Human Language Technology Conference*.

Kristi Seymore, Andrew McCallum, and Ronald Rosenfeld. 1999. Learning hidden markov model structure for information extraction. In *AAAI-99 Workshop on Machine Learning for Information Extraction*.