ABSTRACT
The area of automatic text summarization is becoming more and more important as the amount of information that is freely available skyrockets. This paper describes a novel system of representation for the ideas present in a text or texts. This system is based on syntactic and semantic relationships between words. The representation is used within a program entitled PARE (Pruner and Redundancy Eliminator) to produce a summary of one or more documents. PARE also incorporates a method of pronominal anaphora resolution to limit ambiguity in the summary.

1. INTRODUCTION
It has been said that, “The main point of writing a text ... is to convey ideas, information, or knowledge to the reader” [3]. Even when ideas are not being actively obfuscated, it is often difficult for us to determine exactly what the author wants the reader to understand. It should be no surprise that to create an automated system to do so is similarly challenging. Recent work in the field ([2], [4]) shows that this is still an active field of research.

2. HISTORY AND MOTIVATION
The name of our program, PARE (Pruner and Redundancy Eliminator), is something of an artifact. When work on this project began, a survey of the field suggested that attempting to pick out specific important sentences from within a document or set of documents was a fairly standard method for deriving summaries. The first generation of PARE was an attempt to approach a similar solution from a different direction. The initial idea was that it might be possible to select and remove the unimportant phrases leaving the central ideas, which would then stand as a summary.

Thus, the reductive quality of the program was well-captured by the name. Pruner was used because the first step in producing summaries involved a word-level elimination of stock phrases that add little to no meaning to sentences. This included words in such phrases as “therefore,” “as you can see,” and “in point of fact.” Redundancy Eliminator was used because the second step involved identifying sentences and paragraphs whose ideas were already well-represented in the rest of the text and removing them.

Unfortunately, this approach to the production of a summary did not work as well as we had hoped. Empirically, the results of PARE’s original incarnation were less than exciting. But, more importantly, a reexamination of the ideas reveals some important flaws in our thinking at the time. For instance, in reading a document, one finds that often one sentence in a paragraph summarizes the whole fairly well. Of the rest of the sentences in the paragraph, some of them are simply rewordings or recombinations of ideas that have already been presented. The remaining sentences, particularly in a well-written text, tend to extend by small steps the ideas captured in the exemplar. So, while these sentences are not redundant – because they do introduce new ideas – they can generally be left out of a summary without a significant loss of understanding. Because they are not strictly redundant, they were often left in the final summary by the initial version of PARE, making the summary large and unwieldy and obscuring the most important ideas.

One method that is often used in automatic summarization is to simply count the words of the document and assume that frequently occurring words are important to the meaning of the document. A basic word counting method is not without its merits, and it is the basis for at least one significant implementation of automatic summarization. The “AutoSummarize” tool found in recent versions of Microsoft Word uses such a technique. The following is quoted from the Microsoft Word AutoSummarize help file:

AutoSummarize determines key points by analyzing the document and assigning a score to each sentence. Sentences that contain words used frequently in the document are given a higher score. You then choose a percentage of the highest-scoring sentences to display in the summary.

However, summaries using this technique tend to be choppy and often choose sentences of unclear importance. We feel that to perform “good” summarization there must be at least some consideration of semantics, as ideas and meaning are at the heart of a good summary.
3. GENERATING A REPRESENTATION

We begin by borrowing from semantic linguistics the idea of a semantic network. Diagrams of this sort show the ways in which a number of ideas are related. Figure 1 is one form of semantic network that represents the English sentence, “A Stagirite teacher of a Macedonian conqueror of the world is a disciple and an opponent of a philosopher admired by Church Fathers.” What the network represents is the relationship between ideas like is a stagirite and is the teacher of a man who.

3.1 The graph

PARE begins with a notion founded largely on the concept of a semantic network. PARE creates a graph that represents the relationships between single-word ideas like teacher and Stagirite, or philosopher and admired. The graph is a set of nodes, each of which represents all instances of a single word from the document. These nodes are connected by arcs which contain information describing the relationship between the words within the context of the document. These relationships are determined using a link-grammar-based parser, developed at Carnegie Mellon University by Sleator and Temperley, who created the link-based grammatical system [6]. Essentially, in such a scheme, every word possesses a set of “linking requirements,” which determine the contexts in which the word may appear. For example, most prepositions require an object, which will be found after the preposition in the sentence.

The parser finds the most likely set of links between the words that make up a given sentence, and returns a list of word pairings and link types. Figure 2 shows the results of running the parser on the sentence, “Although John was tired, he managed to climb the wall.” This syntactic analysis provides some level of semantic representation in our graph, as we are able to generalize the link types and retain such valuable information as which nouns take which actions and what actions have which objects.

![Figure 2: An example of the output of the link-grammar parser](image)

Presently, PARE understands these relations between word-level ideas: DETR, DOES, HAS, NAME, OBJ, POBJ, PREP, and QUAL. QUAL is read “has the quality,” so that a link-age like dog-(QUAL)-red means that a dog has the quality of redness – stated in English, “The dog is red.” HAS and DOES are each read as they are written, and represent ownership and action, respectively. OBJ links both prepositions and verbs to their objects. PREP links verbs to prepositions, and POBJ links from the verb to the object of the preposition. NAME connects multiple words that are part of one title, and DETR links determiners to nouns.

Describing ideas at this lower, more atomic, level is significantly easier to do, but the ideas created also seem to be a less fully-formed representation of the full meaning of a sentence than that which is captured by a semantic network. Simply knowing how any two words relate does not necessarily give one a complete understanding of the statement as a whole.

3.2 The colorings

To allow us to capture the meanings of larger groups of words, we introduce a way of representing ideas with size larger than two words in our graph: the notion of a coloring. Each node within the graph may be assigned any number of colors. Colors are assigned so that all nodes with the same color are part of the same idea. This allows PARE to tie together groups of elements and represent within its graph much more complex concepts.

A set of words is considered to be an idea if it is a constituent of a sentence. Within the field of syntactic linguistics, constituents are believed to represent those parts of a sentence which might stand independently as full thoughts. This is the foundation of our use of constituents as low-level syntactic representations of ideas within a document. We use the link-grammar parser to generate constituent trees for each sentence. The subtree beginning at each node of the tree is then given a coloring. Every individual node receives a unique coloring as well.

4. IDENTIFYING THE SIGNIFICANT IDEAS

Once we have a useful representation of the meaning for every sentence of a document, the question still remains:

---

1. This sentence and the network diagram are reproduced from the webpage of J. Sowa [7].

---

![Figure 1: An example of a semantic network](image)
4.1 Ranking the Nodes

Because our attempt to represent meaning takes the form of a graph, we look at established methods of evaluating the importance of subsets of a graph to the whole. One such method comes to mind immediately as having a special relevance to the problem at hand. The PageRank algorithm developed by Page et al., and implemented as the heart of their Google search engine [5] concerns itself with exactly the sort of relational and contextual analysis that we need. We have implemented a modified form of the PageRank algorithm for determining the relative importance of nodes in a semantic graph, such as the one used by PARE.

PageRank is used by Google to provide users with a better idea of which pages containing a search phrase are likely to have useful information. Central to the approach is the idea that hypertext links are a way in which web page administrators suggest that they believe there is some value in another web page. However, Page does not want to simply count the number of links into a webpage, as a link from a trusted source such as The New York Times or slashdot.com carries more weight in the minds of web surfers than one from a random personal web page, and this fact should be worked into the system, if possible.

The PageRank system assigns to each node in a graph a ranking, represented by a floating point value. For a given webpage $q$, we may refer to the set $L_q$ of links out of that page. So, for a page $p$, and a set $B$ of all web pages that link to $p$, the rank of $p$ is $R(p)$, which is calculated by the equation

$$R(p) = \sum_{b \in B} \frac{R(b)}{|L_b|}$$

If a page has rank $r$ and number of links $n$, each page to which it links receives $\frac{r}{n}$.

In the PARE system, this equation is more complex, since we do not assume that all linkages are equal. Instead of $\frac{r}{n}$, a value is passed between two nodes that depends on the type of link that exists between the nodes and the direction in which the link moves. Suppose a graph contains words X, Y, and Z, with linkages X-(DOES)--Y and X-(QUAL)--Z, and also suppose that X has rank 64. We have chosen values for PARE so that a does-out link has value 6 and a qual-out link has value 4. Thus, X has links with total value 10. So, X passes to Y ($\frac{64 \times 6}{10} = 38.4$), or .384, and to Z ($\frac{64 \times 4}{10} = 25.6$). These link values are chosen intuitively — more experimentation is necessary to determine if these numbers are optimal.

This process is recursively defined and is evaluated through repeated iterations until the values converge. Generation of the rankings is implemented using matrices, in a manner similar to that described in the PageRank paper [5]. Their algorithm is as follows:

Let $R$ be a vector of rankings over nodes, and let $A$ be a matrix with rows and columns corresponding to nodes, where $A_{uv}$ is the relationship between $R_u$ and $R_v$, as described above. As pointed out by Page, $S$ may be “almost any vector over [nodes].”

$$R_0 \leftarrow S$$

Loop:

$$R_{i+1} \leftarrow AR_i$$

$$d \leftarrow ||R_i||_1 - ||R_{i+1}||_1$$

$$R_{i+1} \leftarrow R_{i+1} + dE$$

$$\sigma \leftarrow ||R_{i+1} - R_i||_1$$

while $\sigma > \epsilon$

PARE’s algorithm differs from Page’s in more ways than just the adjustable link values. In the original algorithm, $E$ is a vector over nodes, which Page uses to combat the situation dubbed a “rank sink,” in which two nodes link between one another but to nothing else. The vector also serves to mimic a random walk on web pages, emphasizing the egalitarian nature of the Internet, where every web page has some value and probability of being visited based solely on its existence. Because our nodes are more densely linked, and because there isn’t the same sense that randomness is valid or that all nodes have some inherent value, it does not seem clear that $E$ needs to be used. We still create a vector of zeros in $E$, however, because the secondary use of $E$ (as a way of creating rankings personalized about certain nodes) seems like it might be useful in the future, for instance for giving words found in a headline more weight.

It may not be immediately clear that the PageRank algorithm is applicable in this situation. To generalize from the relative popularity of web pages to the relative importance of ideas in a document might sound like a leap. However, it might be argued that PageRank is, in fact, ranking the most important ideas of the internet. Further, the graph contains a significant amount of specific semantic information pertaining to the relationships between word-level ideas. This information is in a form that can be explicitly exploited by the modified PageRank algorithm.

The method is, in some sense, related to more naive word counting, but has several advantages. One is that we don’t have to make an explicit list of stop words, such as the, a, and for. Instead, PARE assigns little value to words on left side of a link such as X-(DETR)--Y or the right side of Y-(PREP)--Z. Similarly, this system easily identifies the principle actors in a document by placing significant weight on the left side of a relationship like M-(DOES)--N, thereby giving higher rankings to those who perform the most interesting actions. Another advantage is that this method captures the value of words that appear few times but consistently appear in relation to important ideas.

4.2 Ranking the Ideas

Once a ranking for each node has been produced, we generate a ranking for each coloring. This color ranking is simply the average ranking of all nodes with that color. These averages are then sorted, producing a list of all of the ideas in the document, from highest average importance to lowest.

The average-based ranking has the desirable property of fa-
5. RESOLVING AMBIGUITY

One possible problem with obtaining a relational understanding of the importance of ideas in a document is the presence of ambiguous phrases in human language. A common example of ambiguity is found in anaphoric phrases. These are words or phrases that refer to other linguistic units that have already been mentioned in the discourse—a common example is a pronoun. In order to capture the semantic importance of anaphoric phrases we need to determine the word or phrase to which they refer.

Consider words such as he or it. Either of these words could refer to any number of ideas within a single document. It is not conceptually sound to associate all of the ideas related to every instance of the pronoun him with a single node in the graph, for example. Likewise, an important association might be missed because the text’s author wanted to avoid repetition.

For this reason, we consider anaphora resolution an important step in the preprocessing of documents to be summarized. Anaphoric expressions come in many forms such as indefinite noun phrases, definite noun phrases, pronouns, demonstratives, and so forth. We developed an anaphora resolution system in which we focus on pronominal resolution. Our decision was based on the fact that pronouns are the most often encountered type of referring expressions and that they offer the biggest set of constraints for their resolution.

In order to resolve a given pronoun, we compile a list of possible noun phrases to which the pronoun could refer. Each of these noun phrases is called a possible antecedent of the pronoun. Pronouns usually refer to nouns; however, we would seriously decrease the validity of our system if only nouns were considered because in many cases pronouns refer to a complete noun phrase. For example, consider the sentences:

"The angry dog chased me. It barked constantly." If we resolve the pronoun it by determining that it refers to the noun dog, we would miss a very important bit of semantic information—the fact that the dog is angry. We extract the noun phrases from a constituent tree of the sentence produced by the previously mentioned link-grammar-based parser.

Once we determine the possible antecedents of a pronoun, scores are assigned to each of them using several rules.

1. The pronoun must agree in number with all of the proposed antecedents. Any antecedents that do not agree are discarded from further consideration.

2. Each remaining antecedent’s score is incremented by a number inversely proportional to the distance of the antecedent from the pronoun. Pronouns usually refer to noun phrases which occur “nearby” in the document. Therefore, the closest antecedent to the pronoun gets the biggest increment to its score, and as antecedents become more distant from the pronoun the number by which we increment their score decreases.

3. Many theories of reference incorporate the notion that a noun phrase that is mentioned many times in a document is more likely to be the referent of a pronoun than a noun phrase that is mentioned less often. This preference is incorporated in our system by computing the mention count of each proposed antecedent and then adding this number to each antecedent’s total score. The mention count of a noun phrase is the sum of the number of times each significant word in the noun phrase appears in the document before the pronoun we are trying to resolve has been mentioned.

4. A pronoun occasionally refers to the same noun phrase that the previous pronoun of the same type refers to. For example, consider the sentences: "John is a good student. He prepares well for every class. He always submits his homework on time." In this example both instances of the pronoun he refer to the noun phrase John. Whenever a pronoun is resolved, the previous pronoun of the same type is found, the noun phrase to which it refers is located, and its score is increased.

Finally, the antecedent with the best score is indicated as the referent of the pronoun being resolved. Currently our system resolves pronouns with about 60% accuracy. We should note that none of the ideas presented here for pronominal resolution are novel to this system. For example, most are discussed in [1]. However, we feel that due to our word graph representation, an accurate method of anaphora resolution is even more critical than usual, and this will be a large focus of future work.

6. AN EXAMPLE

One of the more significant hurdles in the attempt to produce a quality summary automatically is that one must answer the question, “What is a quality summary?” The task of a summarizer is twofold: reduce the word count of the original document, and provide a good amount of information. Doing either of these in and of itself is trivial, as we could simply take every 20th word and reduce the size of the document drastically, or call the document itself a summary and lose zero content. Such policies aside, it is hard to provide quantitatively telling results in the middle ground without extensive testing with human subjects, something that time constraints forced us to omit. While we are working on developing a consistent, useful system for testing the value of our results as summaries, we feel intuitively that we are getting good results.

We provide here an example of our system in practice. Below is the result of running PARE on a July 27, 2003 New
6.1 PARE summary

- I think we must and give this process a chance see how it unfolds Mr Mofaz told Israel radio
- in another development Israel removed and 10 major West Bank checkpoints other barricades that had Palestinian movements blocked limited greatly since they were imposed shortly after the Palestinian uprising began in September 2000
- the Middle East peace plan does not require specifically Israel to release any of Palestinian prisoners the roughly 6,000 it is holding
- but which the rightist Israeli government had said that it would not free members of and Hamas Islamic Jihad groups that have declared a truce oppose peace negotiations with Israel
- but Israel said the barriers inside the West Bank were necessary and to monitor Palestinian movements at multiple points to channel the flow of traffic before would-be attackers approached Israeli towns
- cabinet Mr Sharon’s voted 14 to 9 today to free about 100 jailed members of and Hamas Islamic Jihad the two Palestinian factions that have accounted for most of the suicide bombings against Israel
- the Bush administration moved up the meeting between and Mr Sharon the president to Tuesday with the goal of nudging both sides forward on the peace plan begun formally last month
- and Palestinian leaders factions said the Israeli action fell far short of their demand for the release of all prisoners
- Palestinians and had claimed these checkpoints were punitive largely did not enhance security Israel’s
- news conference in Friday’s with Mr Abbas Bush President was critical of an Israeli security fence that cuts into the West Bank

6.2 Microsoft Word summary

Israeli cabinet agrees to free 100 militants and lift roadblocks.

Prime Minister Ariel Sharon’s government voted today to free about 100 Islamic militants, and the Israeli military lifted several roadblocks in the West Bank. Mr. Sharon’s cabinet voted 14 to 9 today to free about 100 jailed members of Hamas and Islamic Jihad, the two Palestinian factions that have accounted for most of the suicide bombings against Israel.

The Middle East peace plan does not specifically require Israel to release any of the roughly 6,000 Palestinian prisoners it is holding. In Friday’s news conference with Mr. Abbas, President Bush was critical of an Israeli security fence that cuts into the West Bank. He also called on Israel to stop developing Jewish settlements in the West Bank and the Gaza Strip.

The checkpoints dismantled in the past few days had tended to limit Palestinian movement from one Palestinian town to another inside the West Bank.

Israel still has dozens of West Bank checkpoints.

A senior Israeli military commander said Israel was taking an "enormous risk" by removing the barriers.

6.3 Original document

Israeli cabinet agrees to free 100 militants and lift roadblocks.

Prime Minister Ariel Sharon’s government voted today to free about 100 Islamic militants, and the Israeli military lifted several roadblocks in the West Bank. The moves came just hours before Mr. Sharon flew to the United States, where he could face rare differences with President Bush on Middle East peacemaking.

The Israeli actions appeared timed to show that Mr. Sharon’s government was improving conditions for Palestinians, and took place two days after the Palestinian prime minister, Mahmoud Abbas, listed a range of Palestinian demands in his White House visit with President Bush.

The Bush administration moved up the meeting between Mr. Sharon and the president to Tuesday, with the goal of nudging both sides forward on the peace plan formally begun last month. While violence is down sharply, the two sides have taken only limited steps required under the initiative, known as the road map.

Mr. Sharon’s cabinet voted 14 to 9 today to free about 100 jailed members of Hamas and Islamic Jihad, the two Palestinian factions that have accounted for most of the suicide bombings against Israel.

Shaul Mofaz, Israel’s defense minister, voted for the releases, describing the decision as a "hastily dilemma" but necessary for the peace effort.

"I think we must give this process a chance and see how it unfolds," Mr. Mofaz told Israel radio. "It is still a bit early to say, but when examining things over the past month, I can say, with the utmost caution and considerable doubt, that the general direction is positive."

The rightist Israeli government had said that it would not free members of Hamas and Islamic Jihad, groups that have declared a truce, but which oppose peace negotiations with Israel.

The Middle East peace plan does not specifically require Israel to release any of the roughly 6,000 Palestinian prisoners
it is holding. Mr. Abbas, however, has made the issue a top priority. With today's vote, Israel is now expected to release a total of 500 to 600 prisoners in the coming days.

The Israeli authorities emphasized that none of the Palestinians being freed had "blood on his hands." They have either been convicted of relatively minor crimes or have been held without charge.

Palestinian leaders and factions said the Israeli action fell far short of their demand for the release of all prisoners.

"This decision is not enough," said Ismail Haniya, a Hamas leader in the Gaza Strip. "This is only public relations for Sharon before he meets Bush."

Mr. Sharon left for the United States shortly after the Cabinet vote. He has been a frequent White House visitor since he came to power more than two years ago, and the Bush administration has been strongly supportive of Israel in the fighting with the Palestinians.

During this visit, Mr. Sharon planned to emphasize that the Palestinian leadership had failed to arrest and disarm Palestinian factions as required under the road map, Israeli officials said.

Mr. Sharon may also raise other issues in contention. In Friday's news conference with Mr. Abbas, President Bush was critical of an Israeli security fence that cuts into the West Bank. He also called on Israel to stop developing Jewish settlements in the West Bank and the Gaza Strip.

In another development, Israel removed 10 major West Bank checkpoints and other barricades that had blocked or greatly limited Palestinian movements since they were imposed shortly after the Palestinian uprising began in September 2000.

The checkpoints dismantled in the past few days had tended to limit Palestinian movement from one Palestinian town to another inside the West Bank.

Palestinians had claimed these checkpoints were largely punitive and did not enhance Israel's security.

But Israel said the barriers inside the West Bank were necessary to monitor Palestinian movements at multiple points and to channel the flow of traffic before would-be attackers approached Israeli towns. Israel still has dozens of West Bank checkpoints.

A senior Israeli military commander said Israel was taking an "enormous risk" by removing the barriers. He noted that troops continued to prevent planned attacks and that a bomb belt fashioned for a suicide bomber was found this week near the West Bank city of Nablus.

The commander said the military was encouraged by the overall reduction in violence. In addition, the first leg of the security fence, which runs along the northern part of the West Bank, is nearing completion despite the American complaints about its route.

The northern West Bank has been the main launching pad for suicide bombers, and the fence will make it more difficult—but not impossible—for bombers from these areas to reach Israeli towns, the commander said.

7. FUTURE WORK
The obvious shortcoming to someone using PARE is the complete lack of attention to gloss. The ideas are kept in a form that fits well with the graph abstraction, but which does not maintain the original form of the sentences. This means that when the ideas are printed in the summary they often appear slightly mangled—something that is clear from section 6.1. One area with which this project will concern itself in the future is producing friendlier summaries.

One important addition to the PARE system that should help its performance is a stronger system for handling synonyms. A system that claims to capture meaning at a deeper level than word counting must consider synonyms as equivalent words, something that PARE does not currently do. We will also continue work on the anaphora resolution section of the program, to improve its accuracy.

Also, the algorithm for moving larger ideas up in the final list should be more fully investigated, as it puts too much implicit value on simply larger phrases. Some testing is necessary to determine exactly what the best way of providing context without adding useless words would be.

The link-grammar parser utilizes a vast array of different forms of linkages that may exist between words. At the moment, PARE uses about 80% of the links that appear in a given document, among those the most common and those we think are most telling. We may be missing semantic information, however, so we hope to utilize all 100% of the links soon.

Finally, we may consider a move away from specific words in the graph. While the "one word, one node" model is straightforward and easy to understand, it may not capture the fullest semantic value of the sentence or document. We might create a fuller representation of the meaning of a document if, rather than words, our nodes contained semantic forms such as those built into WordNet.

8. SUMMARY
We have presented PARE, an automatic text summarizer. Using a word graph reminiscent of a semantic network, we have developed a ranking mechanism similar to that used by the Google search engine to rank the most important ideas of the document. These ideas are used to generate a summary of the document. While there is room for improvement, we feel that the results are quite good.

9. REFERENCES


