Generic Text Summarization using WordNet

Abstract

This paper presents a WordNet based approach to text summarization. The document to be summarized is used to extract a "relevant" sub-graph from the WordNet graph. Weights are assigned to each node of this sub-graph using a strategy similar to the Google Page-ranking algorithm. These weights capture the relevance of the respective synsets with respect to the whole document. A matrix in which each row repesents a sentence and each column a node of the sub-graph (i.e., a synset) is created. Principal Component Analysis is performed on this matrix to help extract the sentences for the summary. Our approach is generic unlike most previous approaches which address specific genres of documents like news articles *and* biographies. *Testing our system on the standard DUC2000 extracts shows that our results are promising and comparable to existing summarizers.*

Keywords : Summarization, Principal Component Analysis, WordNet, Synset Ranking

1 Introduction

Text summarization is "the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks)" [Man02]. Text summarization finds varied applications in today's world. Some notable ones are: search engine hit summarization (summarizing the information in a hit list retrieved by some search engine); physicians' aids (to summarize and compare the recommended treatments for a patient); generating the blurb of a book ; and so on. Building automated summarizers can be very helpful in many such applications and saves a lot of manual work. An extract is a summary containing only sentences from the text and involves no natural language generation. Given a piece of text, our aim is to select the most "representative" sentences which will form the summary.

A good summary should ideally have the following features

1. Relevance to the text

2. Informativeness

3. Conciseness

1.1 Related work

In this section we cite relevant past literature, that uses various summarization techniques. We then delve into a few of the past approaches in slightly greater detail to motivate and show the uniqueness of our approach, in contrast to the others.

The most common and recent text summarization techniques use either *statistical approaches*, for example [ST98], [BM00], [K.96], [CJ98], [NT01]; or *linguistic techniques*, for example [KJ95a], [RD98], [Nak00], [D.97]; or some kind of a linear combination of these: [JMVJ99], [Man02] and [BR97]. Our algorithm is markedly different from each of these and tries to capture the semantics of the text (refer to section 3).

There has also been a lot of work on text summarization using various kinds of *supervised learning* techniques. Such a technique for summarization was first suggested in [KJ95b]. This type of machine learning approach was studied and worked on in great detail, for example [TS97], [MI98], [CW00] and [BA00]; and [AG02] proposed a *semi-supervised learning* technique for text summarization, where some labeled data are used together with a greater amount of unlabeled data. With the aid of a few particaular case studies, we now motivate our algorithm.

[Zha02] developed an algorithm for summarization based on a mutual reinforcement principle where keyphrases and sentences are ranked based on their saliency scores. A sentence is given a high saliency score if it contains many terms which themselves have a high saliency score. A term is given a high saliency score if it appears in many sentences which have a high saliency score. This is how the mutual reinforcement takes place between sentences and terms. In this approach a sentence is ranked based on terms in the sentence. However, a sentence is not ranked based on some measure with respect to the entire text document. In our approach, a sentence is given weights based on synsets in a part of the WordNet graph that we construct and each synset gets a rank (or a measure of importance) based on the whole graph. This is how the context and content of the whole document reflects in each sentence's rank, and this could be one reason of our algorithm performing well.

[JMVJ99] studies news article summarization and uses statistical and linguistic features to rank sentences in the document. Statistical features are the tf-idf measures, while linguistic features are observed features in a large collection of news articles whose presence in a sentence makes it more probable than not of being a summary sentence. For example, they observed that sentences containing anaphoric references like "these" and "this" do not usually appear in the summary, while sentences containing days of the week appear in the summaries more often than not. These features have been observed specifically for news articles and hence the method is not generic. Also there are always exceptions to the above rules which degrade the quality of the generated summary.

[Man02] has a study of various text summarization techniques. These include:

- location of a term in the document, for example being in the title, beginning of a paragraph or in the conclusion.
- presence of statistically salient terms (using tf-idf measures)

- presence of cue phrases like “on the whole”, “in summary”.
- connectivity of text units based on proximity, repetition, etc.

It is then suggested to take a weighted average of these features to assign a measure of importance to a text unit, for example, a sentence. The weights can be trained from a corpus or be chosen according to the genre of documents being summarized.

We note that none of the above approaches to text summarization selects sentences based on the semantic content of the sentence and the relative importance of the content to the semantics of the text. Our algorithm attempts to do just that. Moreover our algorithm is for generic text summarization unlike some algorithms, for example query-relevant summarization ([BM00], [CY99]), summarization of online material ([BA00] for webpages, [RD98] for summarization of multpile online sources, [RF00] for search engine hit summarization) or news article summarization ([JMVJ99]).

1.2 A brief on the WordNet

WordNet [Fel98] is an online lexical reference system in which English nouns, verbs, adjectives and adverbs are organized into synonym sets or *synsets*, each representing one underlying lexical concept. Noun synsets are related to each other through *hypernymy* (generalization), *hyponymy* (specialization), *holonymy* (whole of) and *meronymy* (part of) relations. Of these, (*hypernymy*, *hyponymy*) and (*meronymy*,*holonymy*) are complementary pairs.

The verb and adjective synsets are very sparsely connected with each other. No relation is available between noun and verb synsets. However, 4500 adjective synsets are related to noun synsets with *pertainyms* (pertaining to) and *attra* (attributed with) relations.

Figure§1 shows that the synset { dog, domestic_dog, canis_familiaris} has a hyponymy link to { corgi, welshcorgi} and meronymy link to { flag} (“a conspicuously marked or shaped tail”). While the hyponymy link helps us answer the question “A corgi is a kind of what?”, the meronymy connection here is perhaps more confusing than useful: this sense of *flag* is rare.

1.3 Our approach

We use WordNet to understand the links between different parts of the document; Subsequently extract the portion of the WordNet graph which is most relevant and contains the main ideas present in the document. We do this by starting from the words in the document as the leaves and traverse the WordNet links upward towards synsets that are more general at each level. This is explained in greater detail in Section 2. The idea of first getting a global view of the whole document, even before beginning to rank sentences is what differentiates our approach from the rest. After obtaining the relevant sub-graph we rank its nodes (synsets of WordNet) with an algorithm explained in Section 2 and use them to compute the importance of individual sentences.

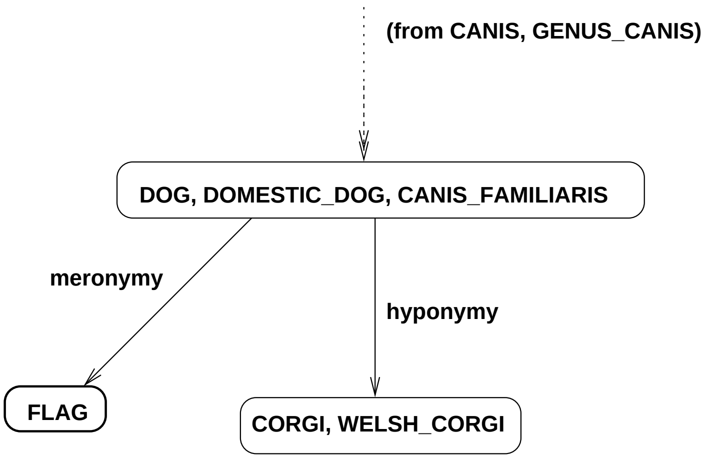

Figure 1: Illustration of WordNet relations.

This idea of getting an overall picture of the document before picking sentences makes our approach generic (not necessarily tailored to give good results only on a specific class of documents).

The last, and the most crucial phase of our approach is to actually pick out sentences that will form the summary. It is important that we do not pick too many sentences with a similar meaning and instead only pick those sentences which represent the text to the maximum extent possible. This is where, we believe, our approach captures human thinking. It is natural for a human to identify very similar sentences from the text and pick not too many of them.

1.4 Roadmap of the paper

Section 2 describes our text summarization algorithm in detail. In section 3, we provide the intuition behind our algorithm by outlining the semantic interpretation of individual steps of the algorithm. Section 4 explains our algorithm through case studies and run through of the algorithm. In section 5, we give our experimental results. In section 6, we give a few concluding remarks. Section 7 talks about possible directions for future work. Finally, the appendix gives four complete examples of single-document summarizations, and their comparison with Copernicus[Cop].

2 Text Summarization Algorithm

The heart of our algorithm is the computation of global semantic information which captures the overall meaning of the text using WordNet.

The algorithm has the following five steps:

1. **Preprocessing of the text:**
    - Break the text into sentences.
    - Apply part of speech tagging to the words in the text. This is essential to pick the correct meaning of the word in WordNet. Hence if the word "pant" is used in the text as a verb, we will not associate it with a form of clothing.
    - Identify *collocations* in the text. A collocation is a group of commonly co-occuring words, for example, "miles per hour". Identifying collocations helps in capturing the meaning of the text better than that of the individual words (just like any idiom). "Miles per hour" would be *closer* to the synset representing "speed" than the three words individually. The WordNet has senses for these collocations, by identifying those set of words as a whole in a synset and not each of them in different synsets. We look for known collocations in the text, and if found replace the independent tagged words, with the collocation as a whole (and hence its synset). Thus we are able to grasp better the sense in which those set of words are used and thereby narrow down the wide spectrum of possibilities.
    - Remove *stop words* like "the", "him" and "had" which do not contribute to understanding the main ideas present in the text.

    The sequence of the above operations is important since we must identify collocations before removing stop words as many stop words often form part of collocations.

2. **Constructing sub-graph from WordNet:** We find the portion of the WordNet graph which is *relevant* to our text, i.e. those synsets whose meaning occurs in the text.

    First we mark all the words and collocations in the WordNet graph which are present in our text. Since the same word may occur in WordNet multiple times under different parts of speech, the part-of-speech tagging helps in narrowing the possible choice of occurance to use, even though it does not determine the choice completely. Now, we traverse the generalisation edges of the WordNet graph starting from these marked words, and keep marking all the synsets which are reachable from the marked words. We do a breadth-first search and traverse the graph only till a fixed depth, as the meanings of synsets become too general to be considered thereafter. Finally, we construct a graph G containing only the marked words and marked synsets as nodes and the generalization links between them as the edges.

3. **Synset Ranking:** The basic motivation of this step is to rank the synsets based on their relevance to the text. So, if lots of words in the text correspond to the same synset, that synset or 'meaning' is more relevant to the text, and thus, it must get a higher rank. This idea has been

borrowed from [RB03], which details the use of WordNet Synsets as a mode of text representation.

We construct a row vector R with an entry corresponding to each node of the graph. The i^{th} entry of R denotes the rank or "importance" of the i^{th} node in the graph, assuming the nodes are numbered from 0 to $n-1$, where n is the number of nodes. We also construct a square matrix A such that

$$\begin{aligned} A[i,j] &= \frac{1}{num\_predecessors(j)}, \text{ if } j \text{ is a child of } i \\ & \quad 0 \qquad\qquad otherwise \end{aligned}$$

where i and j are nodes in the graph G.

Each entry in R is initialized to $\frac{1}{\sqrt{n}}$ Then, we keep applying the following transformation to R until the change in modulus of R becomes less than a fixed tolerance value.

$$R_{new} = \frac{R_{old} * A}{|R_{old} * A|}$$

After each iteration of the above transformation, we add a positive bias to the rank of the nodes corresponding to the words present in the text. This ensures that the rank of the words do not become too low.

4. **Sentence Selection:** The synset ranking algorithm gives us information about which synsets or meanings are more relavant to the text, the relevance being reflected in the rank of the synset. Now, using this information, we have to select sentences from the text based on their relevance to the text keeping in mind that two sentences with a similar semantic content are not both selected.

   We construct a matrix M with m rows and n columns, where m is the number of sentences and n is the number of nodes in G. Initially each M_{ij} is 0. Now, for each sentence S_i, we traverse the graph G starting with the words present in S_i and following the generalisation edges similar to step 2. We find out the set of reachable synsets SY_i. Now, for each synset $sy_{ij} \in SY_i$, we set $M[S_i][sy_{ij}]$ to the rank of sy_{ij} calculated in step 3.

   Now, for a sentence S_i and synset sy_j, we have

$$\begin{aligned} M[S_i][sy_j] &= 0 \quad \text{if } sy_j \text{ is not reachable from } S_i \\ &= R[sy_j] \quad \text{otherwise} \end{aligned}$$

   Intuitively, each row of the matrix represents the meaning captured by the corresponding sentence out of the meaning of the whole text.
   If a synset is reachable from the words of a sentence, then the meaning of that synset is related to that sentence. The row corresponding to a

sentence has non zero entries corresponding to only those synsets which are reachable from it, or whose meaning can be implied by the sentence.

Principal Component Analysis(PCA): We apply PCA on the matrix M and get a set of principal components or eigenvectors. The eigenvalue of each eigenvector is a measure of the relevance of that eigenvector to the meaning of the text, because a higher eigenvalue reflects a higher number of highly ranked nodes along the eigenvector. We sort the set of eigenvectors obtained according to their eigenvalues which also gives their order of relevance to the text. Now, for each eigenvector, we find its projection on all the sentences and select the sentences with the highest projection. We select $n_{numselect}$ top sentences with high projection values in the summary, where $n_{numselect}$ is proportional to the eigenvalue of the eigenvector. If number of sentences to be selected is N, then $n_{numselect}$ corresponding to eigenvector i is $n_{numselect} = (\lambda_i / \sum_j (\lambda_j)) * N$ where λ_i is the eigenvalue corresponding to the ith eigenvector.

The sentences we thus obtain are in order of their relevance to the text. A very notable feature of this method is that we avoid selecting too many sentences with similar semantic content. This occurs since sentences similar in meaning will have a high projection on them from the same eigenvector and we choose only few sentences with highest projection and then look at the next eigenvector. (In case of a tie we choose the sentence with smaller length to make the summary concise). This is the sentence that best represents the meaning contained in the eigenvector. Also, because the eigenvectors are mutually orthogonal to each other in the sentence-synset space, two eigenvectors will not represent a similar meaning or semantic content.

5. **Final filtering**

   The last stage of our algorithm involves the application of simple heuristics to filter out the sentences which have undefined references. These sentences generally do not contribute much to the summary. The heuristics applied are:

   - Removing sentences which contain the words "It","She","He" etc. at the beginning.
   - Removing sentences which begin with quotes.

   The reason for the first heuristics is that these sentences contain undefined references as in "*She was going home.*" and these references cannot be satisfied with the sentences alone i.e. they are bound to nouns outside the sentence.

   Hence it is reasonable to ignore them.

3 Semantic Interpretation of Computational steps

The first step of the algorithm is pre-processing of the text. The requirement of this has already been mentioned above. Firstly, the part of speech tagging is very important to ascertain the right synset of the word in the document. There could still be ambiguity after part of speech tagging, but it at least helps narrow down the choice like the example of "pant" mentioned above. Then, one has to remove stop words like "the", "he" etc. as they only add noise. In the next step of synset ranking, we would otherwise end up with a lot of undesired edges and would thus make it difficult for us to guage the relative importance of the synsets.

The second step is synset ranking. Here, we start with the words present in the text and then draw edges moving upward to synsets they relate to. So, even if different synonyms of a particularly synset is present in the document, there will be many incoming edges and therefore that synset will get a higher weight. Also, this helps in word sense disambiguation. Take for example the word bank. Now if in the text there are other words like shore, sea etc. then these shall relate and connect to the bank synset relating to shore. And there will be fewer incoming edges to the financial bank. On the other hand, if there were words like money, finance, institue etc. then it is more likely that the document has to do with the bank which means financial institution.

Once the synset ranking has been done, sentence selection is done via PCA. As mentioned above, we have a matrix with sentences as rows and synsets as columns. Call this matrix A. This matrix has entries as the importance of synsets for a sentence. So if a synset does not relate to a sentence the corrsponding entry will have zero. And if a sentence relates to a synset twice, then the entry in the matrix will be twice as compared to a sentence that related once. Semantically, this matrix therefore represens which sentences have which synsets or the other way round, which synsets are reflected in which sentences.

In effect, what has been done so far is to identify regions of influence of all the eigen-vectors. Then from each of these similarity islands, representative sentences are chosen. Each island is a set of sentences that closley relate to an eigen vector.

Then on performing AA^t we get a square matrix with as many rows/columns as sentences. This matrix gives us the "cosine similarity" between these sentences. Now, consider an n dimensional graph where n is the number of rows here. And then plot the columns as points on this graph. Now what does each point mean? A particular point shows the correlation of that sentence with various other sentences. Or a particular coordinate of a sentence shows its similarity with that particular sentence. The most prominent eigenvector of this square matrix, in terms of the graph will be that direction which has many points on the vector. This vector represents a hypothetical sentence, which is closest to all sentences in the document. So, now the sentence which is closest to this vector (from among the n) is picked as the first sentence of the summary. This is what we want, because it is this sentence which best represents all the sentences in the document! The same is then done for the remaining eigenvectors also.

4 Illustration of the algorithm

We generated a summary for the following text document. This summary has been generated without the final filtering to illustrate the effectiveness of our approach. The document is:

Bombay is a busy city. Thousands of people come here every day in search of livelihood. No other city in India offers so much of employment opportunity. Since 1970 the population in the city has been steadily hovering around one crore. Next to Bombay come Calcutta, Delhi and Chennai. But, none of these have the dynamism and the charisma that is characteristic of Bombay. It is truly the international city of the country, truly cosmopolitan, truly multi ethnic. No city other than Singapore in South Asia surpasses Bombay in business volume. Probably Dubai comes close, but a very distant close. Indeed Bombay is a very active place.

The summary generated by our approach is quite good:

Bombay is a busy city. Thousands of people come here every day in search of livelihood. No other city in India offers so much of employment opportunity. Next to Bombay, come Calcutta, Delhi and Chennai. But none of these have the dynamism and charisma that is characteristic of Bombay. Indeed Bombay is a very active place .

There are about two sentences that describe how busy the city is: "Bombay is a busy city", and "Indeed Bombay is a very active place", and rightly, one of them is picked, as the first sentence. The other is the last sentence which is very neat as the document is essentially about this aspect of Bombay. The sentence in the summary, "Thousands of people come here every day in search of livelihood." is also suitable for the summary. Overall, it can be seen that the summary does indeed read well, and also represent the text as a whole. If a human being had to summarize the above text, s/he could not have done much better. In the two sentences "Next to Bombay, come Calcutta, Delhi and Chennai. But none of these have the dynamism and charisma that is characteristic of Bombay." we even find correct pronoun binding. And these two also lend to the summary reading complete and clean.

We also ran our summarizer on the following two "two sentence" texts. Even though these are very simple examples they bring out two very important points.

John is a bright boy. He has a stupid brain.

This is of course a very meaningless document, but it nevertheless illustrates a very important point. The summary generated by our approach, for the above text, was again the two line document itself! This is explained by the fact that the two sentences contradict each other, and hence, both sentences will have high components along different eigenvectors. So, each of these sentences would

	Copernicus	Ours
DUC-200	0.7864	0.7402
DUC-400	0.8402	0.7875

Table 1: Average Similarity Measure

be most important along their respective eigenvectors, and both are picked. Moreover, from a linguistic point of view, if only one of them was picked, the resulting summary would be grossly inaccurate. In order to summarize the text, and retain all important information, the only suitable extract is all of it.

The second example we tried our approach on is:

John is a bright boy. He is a brilliant lad.

In this case, both the sentences mean more or less the same. Notice that since it is a two sentence text, both sentences are very important, in the sense that both capture the essence of the text. Since the two sentences are similar, in the matrix of sentence versus synset, both these will get similar ranks for the various synsets. Hence, the eigenvectors will have almost an equal component along both these sentences. Therefore, on using a threshold between succesive eigenvectors, only one of them will be picked. Logically also, since both have the same import, having them in the extract would make one of them redundant. Our summarizer picks only one of them.

There are four examples of larger document summarization in the appendix.

5 Results

We tested our summarizer using the DUC'2000 multi-document summary coprus. To compare our summaries with those of the well known Copernicus summarizer [Cop], we use DUC's human generated summaries as the benchmark. The results are very encouraging. DUC-200 and DUC-400 denote the DUC summaries of length 200 words and 400 words respectively.

Table 1 gives the average similarity between DUC-200 and DUC-400 with the summaries generated by our algorithm and that of Copernicus. The similairy measure used is the cosine similarity.

Figure 2 plots the similarity between our summaries with DUC-200 and that of Copernicus' with DUC-200. Figure 3 plots the similarity between our summaries with DUC-400 and that of Copernicus' with DUC-400. Dotted plots are that of Copernicus' while ours are in dashed.

In the figures the document number is plotted on the $x-axis$ and the cosine similarity on the $y-axis$. The horizontal lines in the plots show the average similarity.

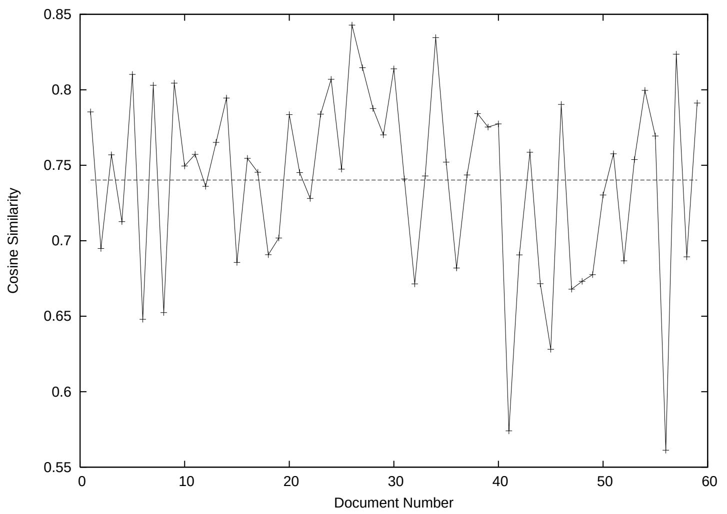

Figure 2: Similarity with DUC-200.

6 Conclusion

There is an ever-increasing need for better automatic text summarization systems with the explosion in the amount of information available. Most existing text summarization systems analyze a text statistically and linguistically, determine important sentences, and generate an extract for it. The linguistic features used are generally specific for a particular kind of document which make existing systems very specialized. We propose an algorithm for a generic text summarizer which selects sentences on the basis of their semantic content and its relevance to the main ideas contained in the text. We use WordNet to abstract the ideas contained in the text so that sentences are selected on the basis of their meaning, and not on the presence of some keywords or frequently-occuring term. Our algorithm compared well with the DUC'2000 data sets and their reference summaries. Eventhough we just concatenated these multi-documents of DUC and ran them as a single document, the results obtained are encouraging. So, on single documents, it is quite possible that ours is even more promising. Even though our average results are slightly worse than that of Copernicus', our algorithm is simple, repeatable and the results can be verified, unlike Copernicus'. Moreover, ours is a novel approach and therefore, building on this (perhaps using some heuristics or otherwise) could improve results drastically. And this approach could be of independent interest.

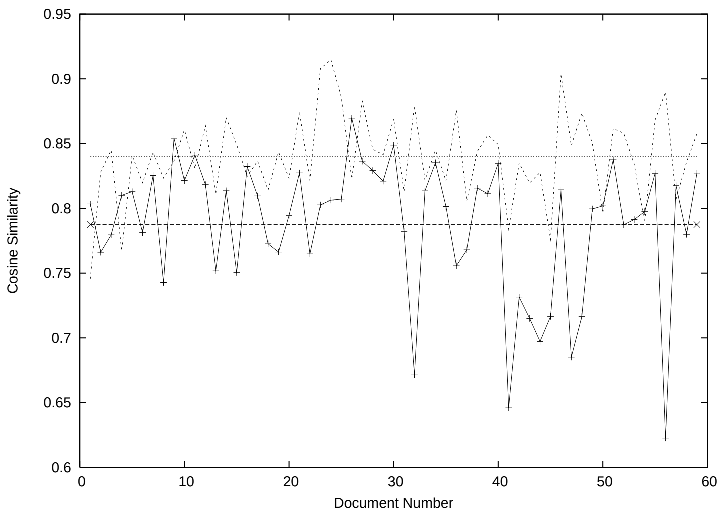

Figure 3: Similarity with DUC-400.

7 Future Work

A number of interesting possibilities remain that we hope to explore in future work. For example,

- *Learn parameters:* The parameters used for generating summaries, eg, weightage given to different parts of speech, depth of the WordNet graph used and number of iterations of the ranking algorithm can be learned given a corpus of documents.

- *Anaphora resolution:* Resolution of pronouns can be used on top of the WordNet approach to get summaries which are more readable and have less dangling anaphors. As an example, if we construct a graph with each node representing each sentence, and an edge from sentence S_i to S_j if S_i has a pronoun referring to a noun in S_j, then we can give higher ranks to sentences with more incoming edges and probably not pick a sentence with an outgoing edge unless the sentence to which it has an edge is also picked.

- *Extend to Other Languages:* Since the approach we have used is not language specific, we can use a lexical resource like WordNet for other languages and extend the summarizer to them.

References

[AG02] Massih-Reza Amini and Patrick Gallinari. The use of unlabeled data to improve supervised learning for text summarization. *Proceedings of the 25th ACM SIGIR*, 2002.

[BA00] Mittal V.O. Berger A.L. A system for summarizing web pages. *Research and Developement in Information Retrieval*, 2000.

[BM00] A. Berger and V. Mittal. Query-relevant summarization using faqs. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000.

[BR97] Elhadad M. Barzilay R. Using lexical chains for text summarization. *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization.*, 1997.

[CJ98] Goldstein J. Carbonell J.G. The use of mmr, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st ACM SIGIR*, 1998.

[Cop] Copernicus summarizer:. *http://www.copernic.com/en/products/summarizer/*.

[CW00] Yang J. Chuang W.T. Extracting sentence segments for text summarization: a machine learning approach. *Proceedings of the 23rd ACM SIGIR.*, 2000.

[CY99] Szpakowicz S. Chali Y., Matwin S. Query-biased text summarization as a question-answering technique. *Proceedings of the AAAI Fall Symposium on Question Answering Systems.*, 1999.

[D.97] Marcu D. From discourse structures to text summaries. *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization.*, 1997.

[Fel98] Christiane (Ed.) Fellbaum. *Wordnet : An Electronic Lexical Database (Language, Speech and Communication).* MIT Press, 1998.

[JMVJ99] Goldstein J., Kantrowitz M., Mittal V., and Carbonell J. Summarizing text documents: Sentence selection and evaluation metrics. *Proceedings SIGIR*, 1999.

[K.96] Zechner K. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. *COLING*, 1996.

[KJ95a] Shaw J. Klavans J.L. Lexical semantics in summarization. *Proceedings of the First Annuwal Workshop of the IFIP working Group for Natural Language Processing and Knowledge Representation*, 1995.

[KJ95b] Chen F. Kupiec J., Pedersen J. A trainable document summarizer. *Proceedings of the 18th ACM SIGIR.*, 1995.

[Man02] Inderjeet Mani. Automatic summarization. *A tutorial presented at ICON*, 2002.

[MI98] Bloedom E. Mani I. Machine learning of generic and user focussed summarization. *Proceedings of the Fifteenth National Conference on AI.*, 1998.

[Nak00] Yoshio Nakao. An algorithm for one-page summarization of a long text based on thematic hierarchy detection. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000.

[NT01] Matsumoto Y. Nomoto T. A new approach to unsupervised text summarization. *Proceedings of the 24th ACM SIGIR*, 2001.

[RB03] G. Ramakrishnan and P. Bhattacharya. Text representation with wordnet synsets. *Eight International Conference on Applications of Natural Language to Information Systems (NLDB2003)*, 2003.

[RD98] McKeown K. Radev D. Generating natural language summaries from multiple online sources. *Computational Linguistics*, 1998.

[RF00] Dragomir Radev and Weiguo Fan. Automatic summarization of search engine hit lists. *Proceedings, ACL Workshop on Recent Advances in NLP and IR*, 2000.

[ST98] Wise B. Strzalkowski T., Wang J. A robust practical text summarization system. *Proceedings of the Fifteenth National Conference on AI*, 1998.

[TS97] Moens M. Teufel S. Sentence extraction as a classification task. *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization.*, 1997.

[Zha02] Hongyuan Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. *SIGIR*, 2002.

A Appendix

Here we present four specific cases of text summarization for single documents, along with their generated summaries using our new approach and that of Copernicus. We have tried to be eclectic in chosing the examples, and present one example where our summarizer outperforms the Copernicus one, and also an example where our summarizer is outperformed by the Copernicus one. We also give one example each of an instance where both our summarizer and the Copernicus perform well and where both perform below par. At the end of each set of document and summaries, we also provide minimal comments regarding the set.

Example1:
What follows is an example where our summarizer produces a better result than the Copernicus summarizer.

Document:
Exotic Kashmir, a tourist paradise of houseboat hotels and Mogul gardens from whose name the English made "cashmere" has become a war zone of separatism and religious enmity. Military jeeps carry senior government officials and civil servants around Srinagar in these dangerous days, as soldiers and paramilitary troopers try to stop a campaign of violence by Moslem separatists in the Kashmir Valley. Intelligence officer M.K. Kaul commands one of the jeeps. At the beginning of a trip through the city of 1 million people, he turned to the four soldiers in the jeep's rear seats and ordered: "Check your weapons. Be ready to fire." Kaul settled into the front seat, a 9mm pistol in his right hand. "These are bad days", he murmured. "These die-hard terrorists can attack from anywhere. To his men, he said: "Shoot them if you see them." On Monday, an around-the-clock curfew was lifted from 5 a.m. until noon and stores opened to crowds of customers. Soldiers in sandbagged bunkers, armed with light machine guns, guarded main intersections. A spokesman for the Jammu-Kashmir state government said the situation was peaceful and "absolutely normal" in Srinagar, the state's largest city. At least 72 people have been killed since the crackdown began Jan. 20 in an effort to curb attacks on government officials and buildings. Most of the dead have been Moslem militants shot by Indian troops. "We are neither fighting a war nor trying to suppress any popular movement", said Jagmohan, governor of Jammu-Kashmir, the only state in predominantly Hindu India with a Moslem majority. "We are fighting a handful of terrorists who are determined to create a law and order problem", said Jagmohan, a Hindu with a reputation for toughness. Other officials in Kashmir say privately the situation is critical, perhaps the most serious threat India has faced in 42 years of independence from Britain. "It looks like everyone here has turned into a fundamentalist with firm belief that the day is not far off when they will become John is a bright boy. He has a stupid brain. independent", a state police officer said. Kashmir was a separate kingdom, ruled by a Hindu maharajah, when Britain relinquished colonial rule on Aug. 15, 1947, and the subcontinent was partitioned on religious lines into India and Pakistan. The Hindu king opted to join India, not Pakistan, and the dispute over Kashmir has exacerbated relations since. Pakistan took the western part of Kashmir and India the eastern. Each accused the other of forcible occupation. They have fought wars over Kashmir and the border question remains unresolved. U.N. observers are stationed on the truce line and the two sides occasionally trade gunfire. Officials will not say how many reinforcements have been sent to the Kashmir Valley since Jan. 20, but reports from various sources indicate 30,000-and 40,000 soldiers and paramilitary troops have been deployed. It is the valley's heavily Moslem population that tilts the sectarian scale in Jammu-Kashmir state. Of the state's nearly 6 million people, 64 percent are Moslems and 32 percent Hindus. Most Hindus live in the southern Jammu area. Srinagar has 150 mosques. "This is the best breeding ground

for Moslem fundamentalism", said a federal official who helps coordinate the new security measures. Nearly every mosque has a loudspeaker to summon the faithful to prayer and, more recently, to exhort listeners to defend their faith and Kashmir. "The residents are constantly bombarded with Islamic sermons and are reminded that fighting a 'jihad' (Islamic holy war) is the most sacred duty of every Moslem", the federal official said, on condition of anonymity. Electricity was cut off in the city Friday, which was both the Moslem sabbath and India's national Republic Day on India's Republic Day, but many mosque loudspeakers continued operating because militants connected them to car batteries. Intelligence reports suggest the government has an advantage because the separatists are not united. The most powerful group is the Jammu-Kashmir Liberation Front, but there are several others, including the Allah Tigers, Hezb-e-Mujahed Hezb-e-Islami. India claims the fundamentalists get money and weapons from Pakistan. The Pakistanis deny it. Jagmohan, the governor, insists the main issue in Kashmir is economic development, not separatism. "The people do not want violence or the so-called independent Kashmir", he said. "The need of the hour is development. Let more tourists come. Let new schools be built. Let new hospitals come up." As he spoke, military officers waited in the next room to discuss a new strategy to crush rebellion in what the tourist brochures call Paradise. Right now, Paradise looks lost, at least temporarily.

Our Summary:
Military jeeps carry senior government officials and civil servants around Srinagar in these dangerous days, as soldiers and paramilitary troopers try to stop a campaign of violence by Moslem separatists in the Kashmir Valley. Kaul settled into the front seat, a 9mm pistol in his right hand. On Monday, an around-the-clock curfew was lifted from 5 a.m. until noon and stores opened to crowds of customers. Soldiers in sandbagged bunkers, armed with light machine guns, guarded main intersections. At least 72 people have been killed since the crackdown began Jan. 20 in an effort to curb attacks on government officials and buildings. Most of the dead have been Moslem militants shot by Indian troops. Other officials in Kashmir say privately the situation is critical, perhaps the most serious threat India has faced in 42 years of independence from Britain. The Hindu king opted to join India. not Pakistan, and the dispute over Kashmir has exacerbated relations since. Pakistan took the western part of Kashmir and India the eastern. They have fought wars over Kashmir and the border question remains unresolved.

Copernicus Summary:
Exotic Kashmir, a tourist paradise of houseboat hotels and Mogul gardens from whose name the English made "cashmere" has become a war zone of separatism and religious enmity. Military jeeps carry senior government officials and civil servants around Srinagar in these dangerous days. as soldiers and paramilitary troopers try to stop a campaign of violence by Moslem separatists in the Kashmir Valley. At the beginning of a trip through the city of 1 million people, he turned to the four soldiers in the jeep's rear seats and ordered: "Check

your weapons. “We are neither fighting a war nor trying to suppress any popular movement”, said Jagmohan, governor of Jammu-Kashmir, the only state in predominantly Hindu India with a Moslem majority. Other officials in Kashmir say privately the situation is critical, perhaps the most serious threat India has faced in 42 years of independence from Britain. “It looks like everyone here has turned into a fundamentalist with firm belief that the day is not far off when they will become independent”, a state police officer said. Kashmir was a separate kingdom, ruled by a Hindu maharajah, when Britain relinquished colonial rule on Aug. 15, 1947, and the subcontinent was partitioned on religious lines into India and Pakistan. Electricity was cut off in the city Friday, which was both the Moslem sabbath and India's national Republic Day on India's Republic Day, but many mosque loudspeakers continued operating because militants connected them to car batteries. India claims the fundamentalists get money and weapons from Pakistan.

Comments:
Clearly, the summary generated by our approach isi quite coherent and readable, and at the same time representative of the essence of the document (though not comprehensive, because of the small size). An interesting observation is that our summary does not pick the very first sentence of the document. It is common knowledge that very often the first sentence is a good sentence for a summary (it can be seen that Copernicus picks it). However, it is noteworthy that in this exception, there is no need to pick the first sentence if the second one is being picked, since the second sentence conveys the meaning of the first sentence, in addition to giving other information. So, the second sentence is a perfect start to the summary. A finer points of critique of the Copernicus summary is the unbound pronoun reference in *"he turned to the four ..."*.

Example 2:
Here we present an example where both our summary and that of Copernicus are reaonably good (representative and concise).

Document:
A violent storm that spun tornadoes across the South and Midwest blew north Thursday, knocking a cafeteria wall down on top of lunching schoolchildren in upstate New York. The two-day toll was at least 27 dead and hundreds injured. The cafeteria cave-in claimed the lives of seven children near Newburgh, N.Y., and deaths also were reported in New York City, New Jersey and Illinois. Seventeen people died Wednesday in Huntsville, Ala., when a tornado packing 250 mph winds struck virtually without warning, laying waste to a wide swath of the city. The storm churned around a collision of record warm air and extremely cold air. The result was tornadoes and unusually severe thunderstorms. By far the worst damage was in Huntsville, where an estimated 1,000 people were left homeless and National Guardsmen stood watch against looting Thursday. No bodies were believed left in the rubble, but portions of the city resembled a bombing target. A shopping center and clusters of apartment buildings were

leveled. Cars were piled atop each other. Utility poles had been snapped like twigs. The National Weather Service said the tornado path had covered between eight and 10 miles and that its wind speeds were as high as 250 mph. “It’s like taking six to 10 city blocks and putting them in a blender and putting it on liquefy”, said rescue worker Bob Caraway. As the storm moved north on Thursday, it knocked over trees and ripped the the roofs off houses in West Virginia, injuring four members of one family, two seriously. The storm also caused flooding in Washington, D.C., blew the windows out of skyscrapers in Philadelphia and destroyed a house in Piscataway, N.J. A truck driver died when his tractor-trailer overturned in high winds on a bridge connecting Elizabeth, N.J., and New York City. A New York City woman died after being hit by a steel beam blown off a water tower. In Wyckoff, N.J., high winds knocked a power line onto a gas line, sparking a fire that forced the evacuation of 20 homes, according to police. No houses were burned. The cafeteria collapse occurred at the East Coldenham Elementary School in Newburgh at about 12:30 p.m., when from 113 to 125 students in first, second and third grades were eating lunch, said Newburgh police Chief John Kulisek. State Police and survivors said they believed the school was hit by a tornado, but the National Weather Service said it could not confirm that until investigators went to the scene. The center section of a cinder-block and glass wall fell in on the cafeteria, showering the pupils with glass and concrete. “You are talking about huge concrete blocks that fell on top of the kids”, state police Lt. Robert Hughes said. Seven children ages 7 and 8 were killed and 18 others were injured, four critically, officials said. In eastern Connecticut, more than 75,000 Northeast Utilities customers lost power in the afternoon, and 18,000 remained without power at 9 p.m., spokeman Jeffery Kotkin said. Tornadoes were reported Wednesday in Mississippi, Alabama, Georgia, South Carolina, North Carolina, Kentucky, Ohio and Indiana. A severe thunderstorm in Illinois claimed one life and caused several million dollars in damage. Altogether, the storms were blamed for at least 27 deaths and more than 500 injuries over the two days. In Alabama, Gov. Guy Hunt said he would ask President Bush to declare Huntsville a disaster area to open the door to millions of dollars in federal relief. Mayor Steve Hettinger said no firm damage estimates were expected until Friday, but he put the number of homeless at 1,000. Hunt assigned nearly 200 Guardsmen to assist in the cleanup and help law officers guard against looting. Police spokeswoman Susan Williamson said there had been minor looting the night of the storm but no arrests. The temperature, which had risen to 73 before the storm Wednesday, plunged into the 30s early Thursday while rescuers worked through the night under the glare of floodlights. The lights, powered by generators, produced an eerie glow in a area that was otherwise largely without power. Tall cranes lifted the heavy rubble, while workers with flashlights picked through the smaller pieces. As daybreak came, people began returning to the scene, some to look for valuables in ravaged cars or businesses, others simply out of curiosity. Many carried cameras or videocassette recorders. Police Maj. Robert Moder said there was no indication that anyone was left in the rubble. “We don’t have any reports of people missing”, he said Thursday morning. Moder said 463 people were injured and more than 150 of those required at least overnight hospital treatment. Authorities had said 19 people were killed but later lowered the death toll by two. All but one of the victims, who ranged in age from 7 to 67, were killed along Airport Boulevard, a busy thoroughfare that seemed to be a guide for the tornado on its deadly march. The other victim perished on the city's outskirts, authorities said. Most of the dead were in apartments and stores that were torn into twisted wreckage, or in cars blown off the pavement. The tornado seemed to catch everyone by surprise. "It was fast", said Lucy Lee Rusk, who moved to a Red Cross shelter after her apartment was battered by debris. "It was like one big pop and that's when everything went." The National Weather Service had issued a severe thunderstorm warning at 3:54 p.m., but the tornado struck with no real warning, said Brenda Page of the service's office. A tornado warning was broadcast at 4:39 p.m., two minutes after the twister was sighted on the ground at the old Huntsville airport. Ms. Page said radar showed a strong storm but not the conditions usually visible during tornadoes. Danny Cooper, director of the Alabama Emergency Management Agency, said three people were seriously injured in an earlier tornado that hit a home in Clay County in eastern Alabama. He said a tornado in northwestern Alabama's Lauderdale County destroyed three homes. Alabama Power Co. reported Thursday that violent winds had battered the switchyard of the Farley Nuclear Power Plant near Dothan on Wednesday night, causing about three gallons of concentrated PCBs to spill. A spokesman said no one was injured, the nuclear plant was not affected and a cleanup of the cancer-causing chemicals was under way.

Our Summary:
A violent storm that spun tornadoes across the South and Midwest blew north Thursday, knocking a cafeteria wall down on top of lunching schoolchildren in upstate New York. The two-day toll was at least 27 dead and hundreds injured. The cafeteria cave-in claimed the lives of seven children near Newburgh, N.Y., and deaths also were reported in New York City, New Jersey and Illinois. Seventeen people died Wednesday in Huntsville, Ala., when a tornado packing 250 mph winds struck virtually without warning, laying waste to a wide swath of the city. The storm churned around a collision of record warm air and extremely cold air. By far the worst damage was in Huntsville, where an estimated 1,000 people were left homeless and National Guardsmen stood watch against looting Thursday. No bodies were believed left in the rubble, but portions of the city resembled a bombing target. The National Weather Service said the tornado path had covered between eight and 10 miles and that its wind speeds were as high as 250 mph. As the storm moved north on Thursday, it knocked over trees and ripped the the roofs off houses in West Virginia, injuring four members of one family, two seriously. The cafeteria collapse occurred at the East Coldenham Elementary School in Newburgh at about 12 : 30 p.m., when from 113 to 125 students in first, second and third grades were eating lunch, said Newburgh police Chief John Kulisek. Hunt assigned nearly 200 Guardsmen to assist in the cleanup and help law officers guard against looting.

Copernicus Summary
A violent storm that spun tornadoes across the South and Midwest blew north Thursday. knocking a cafeteria wall down on top of lunching schoolchildren in upstate New York. The cafeteria cave-in claimed the lives of seven children near Newburgh. N.Y.. and deaths also were reported in New York City. New Jersey and Illinois. Seventeen people died Wednesday in Huntsville. Ala.. when a tornado packing 250 mph winds struck virtually without warning. laying waste to a wide swath of the city. As the storm moved north on Thursday. it knocked over trees and ripped the the roofs off houses in West Virginia. injuring four members of one family. two seriously. The cafeteria collapse occurred at the East Coldenham Elementary School in Newburgh at about 12 : 30 p.m.. when from 113 to 125 students in first. second and third grades were eating lunch. said Newburgh police Chief John Kulisek. State Police and survivors said they believed the school was hit by a tornado, but the National Weather Service said it could not confirm that until investigators went to the scene. In eastern Connecticut, more than 75,000 Northeast Utilities customers lost power in the afternoon, and 18,000 remained without power at 9 p.m., spokeman Jeffery Kotkin said. "We don't have any reports of people missing", he said Thursday morning. He said a tornado in northwestern Alabama's Lauderdale County destroyed three homes.

Comments:
Again, one might observe that both the above summaries do capture the essence of the text and are thus representative. Also both of them are quite readable. Interestingly, there is quite a lot of overlap between the two summaries, as well.

Example 3:
In the following example, the summary generated by our approach is not satisfactory. On the contrary, Copernicus' summary is better.

Document:
Planes loaded with food, medicine and tents are converging on earthquake-stricken Iran from around the world, among them the first private American relief mission to the country in more than a decade. Authorities now are faced with the difficult task of getting supplies into the affected areas. More than 60 aftershocks jolted the quake zone in northwestern Iran on Sunday, triggering landslides along one key relief route. Tehran radio said Sunday that 68 relief aircraft had landed at Tehran's Mehrabad airport during a 24-hour period. Among them was a plane carrying medicine and other supplies from the United States. Another plane carrying U.S. government aid was due today. The shipment that arrived Sunday was the first of its kind to be accepted by Iran since the hostage crisis of 1979-80, when 52 Americans were held at the U.S. Embassy in Tehran. Two countries that also have suffered devastating earthquakes in recent years, the Soviet Union and Mexico, were among those sending assistance to help victims of Thursday's quake. Leaders of Iran's Arab neighbors, including Iraq, which waged an eight-year war with Iran, issued statements of sympathy

and support. Pope John Paul II called on the international community to help earthquake victims, asking for "a real and generous demonstration of solidarity with our brothers. Let us pray that their suffering is alleviated." The Vatican was among the many international donors. Eleven planes from West Germany ferried in 145 tons of supplies plus a field hospital with two medical teams. Nine flights arrived from Turkey. The Spanish Red Cross sent medicines and food; India sent medicine. The Austrian Red Cross shipped tents, blankets and medicine. Bernard Kouchner, the French secretary of state for humanitarian action, arrived in Iran on Sunday to coordinate French assistance and evaluate future needs, Radio France Inter said. The French government has sent a civil defense team of 195 people and 18 search dogs along with at least 24 tons of equipment. The Saudi-based 46-member Organization of the Islamic Conference, an umbrella organization for the world's estimated one billion Moslems, issued an "urgent appeal to the international community and especially Islamic countries and organizations to hasten with emergency aid for the Iranians..." AmeriCares, a private U.S. relief organization based in New Canaan, Conn., flew in the 42 tons pounds of supplies that arrived Sunday. "To be on the ground here in Tehran is a very special feeeling for us", said Stephen M. Johnson, president of AmeriCares. A spokeswoman for the U.S. Agency for International Development said a Red Cross charter flight carrying $225,000 worth of supplies donated by the U.S. government would arrive in Tehran this morning. Iran is believed to have influence with the Shiite Moslem extremists holding 16 Western hostages, including six Americans, in Lebanon. The longest-held is Terry Anderson, chief Middle East correspondent of The Associated Press. He was kidnapped March 16, 1985. The Soviet Union sent in medical teams and supplies from Soviet republics bordering Iran. More than 200 medical personnel arrived from the Soviet Union, which also sent medical supplies, trucks, trailers, ambulances, cranes, buses and water tankers. The December 1988 earthquake in Soviet Armenia left at least 25,000 people dead. Mexico, which lost at least 10,000 people in a 1985 earthquake, is sending food, medical supplies and blankets, officials said Sunday. In Lebanon, the Iranian-backed Hezbollah, or Party of God, said its members were donating 10 percent of their salaries to help the victims. That would amount to about $500,000, according to Hezbollah sources, who spoke on condition of anonymity.

Our Summary:
Planes loaded with food, medicine and tents are converging on earthquake-stricken Iran from around the world. among them the first private American relief mission to the country in more than a decade. Authorities now are faced with the difficult task of getting supplies into the affected areas. More than 60 aftershocks jolted the quake zone in northwestern Iran on Sunday, triggering landslides along one key relief route. Among them was a plane carrying medicine and other supplies from the United States. The Spanish Red Cross sent medicines and food; India sent medicine. Bernard Kouchner, the French secretary of state for humanitarian action, arrived in Iran on Sunday to coordinate French assistance and evaluate future needs, Radio France Inter said.

Copernicus Summary:
Planes loaded with food, medicine and tents are converging on earthquake-stricken Iran from around the world, among them the first private American relief mission to the country in more than a decade. Authorities now are faced with the difficult task of getting supplies into the affected areas. Among them was a plane carrying medicine and other supplies from the United States. The shipment that arrived Sunday was the first of its kind to be accepted by Iran since the hostage crisis of 1979-80, when 52 Americans were held at the U.S. Embassy in Tehran. Two countries that also have suffered devastating earthquakes in recent years, the Soviet Union and Mexico, were among those sending assistance to help victims of Thursday's quake. Pope John Paul II called on the international community to help earthquake victims, asking for "a real and generous demonstration of solidarity with our brothers. Eleven planes from West Germany ferried in 145 tons of supplies plus a field hospital with two medical teams. Bernard Kouchner, the French secretary of state for humanitarian action, arrived in Iran on Sunday to coordinate French assistance and evaluate future needs, Radio France Inter said. The French government has sent a civil defense team of 195 people and 18 search dogs along with at least 24 tons of equipment. Saudi-based 46-member Organization of the Islamic Conference, an umbrella organization for the world's estimated one billion Moslems, issued an "urgent appeal to the international community and especially Islamic countries and organizations to hasten with emergency aid for the Iranians..." AmeriCares, a private U.S. relief organization based in New Canaan, Conn., flew in the 42 tons pounds of supplies that arrived Sunday.

Comments:
The main drawback of our summary above is that it does not comprehensively present the content of the entire document. The Copernicus summary, on the other hand, can be seen to be more complete, though there are problems with this summary as well. For example, in *"Among them was a plane..."* seems to incorrectly refer to the *Authorities* mentioned in the previous sentence of the summary.

Example 4:
As a last example, we present a case where both approaches (our approach and that of Copernicus) fail to produce satisfactory results.

Document:
The international community is starting to look beyond the conflict in the former Yugoslavia, and is becoming increasingly aware of the true magnitude and seriousness of ethnic and political conflicts in the former Soviet Union. The problem is indeed a pressing one. The disputes in the Transcaucasus, central Asia, and elsewhere in the former Soviet Union are causing great suffering for those involved. They are a source of legitimate concern to the Russians, who are worried by clashes close to their borders. They are concerned too about the

safety of ethnic Russians living in those countries where there are conflicts, and about the economic and social consequences of refugees fleeing to Russia (whose potential number is estimated at more than 2m). They believe that the problems in some of the former Soviet republics are either ignored or misunderstood. The Russians fear that if they do nothing, the conflicts will worsen and probably spread. They argue that they have a legitimate concern in trying to resolve conflicts that threaten their security. They ask why, when the international community is engaged in peacemaking and peacekeeping in many of the world's trouble spots, similar efforts should not be undertaken in the hot spots near them. Equally, these disputes are of legitimate concern to Britain, its western partners and the whole international community, both for humanitarian reasons and because such flashpoints risk provoking regional instability. If uncontrolled, they pose a variety of threats to European security, including increased traffic in drugs and arms as well as international terrorism. Crises in Russia's newly independent neighbours may also complicate Russia's relations with the west: unless they are carefully handled, they risk prejudicing the emerging relationship between us. The west is sometimes suspicious of Moscow's attitude. There are no differences about the need to resolve these conflicts: but there are differences over how to go about it. The British government will not underwrite Russian involvement unless it is sure that peacekeeping troops will not turn into occupying forces. Nor are countries that secured their independence when the Soviet Union was dissolved keen that the west should concede to Moscow the same role it played in the Soviet empire. They fear that this would compromise their sovereignty. When they are willing for Russian troops to play a peacekeeping role, they want there to be some form of international monitoring. This has meant that, even though the west lacks the resources to engage directly in those trouble spots, it has been wary of appearing to endorse Russian involvement instead, in case it had imperial tendencies. If we are to fashion a joint approach by Russia and its western partners to the problem of peacekeeping in the former Soviet Union, we must address all these concerns. Clearly Russia has an important role to play in this area. If the Russians want international recognition and more active support from the international community, the west in turn needs guarantees that Russia will not regard recognition of its involvement as giving it carte blanche in the former Soviet Union. Peace-keeping in the former Soviet Union, as indeed peacekeeping elsewhere, must comply with internationally recognised principles. It needs to be looked at on a case by case basis. Respect for the sovereignty of the newly independent countries must remain paramount. We agreed that the principles underlying Russian, CIS or international peacekeeping operations would need to include: strict respect for the sovereignty of the countries involved; an invitation from the government concerned and the consent of the parties to the conflict; commitment to a parallel political peace process; a clear mandate setting out the role of the peacekeeping forces (which should whenever possible be multinational in character); and an exit strategy for the peacekeeping forces deployed. These principles should help stimulate a dialogue and political settlement between the parties to the conflict, focus attention and, where possible, resources on the settlement of conflicts in the former Soviet Union. All this will help enhance Russian peacekeeping practices. Although the Conference for Security and Co-operation in Europe and the UN lack the resources to put large numbers of peacekeepers on the ground themselves, they may be able to suggest measures to enhance transparency and to monitor the peace process. Extending co-operation within the North Atlantic Co-operation Council, and also co-operation between Russia and Nato, might help improve training for Russian peacekeepers. This is an area we shall need to explore further. We also need to put much more emphasis on preventive diplomacy. One of the ways in which we can help prevent conflicts is to strengthen the democratic institutions in the newly independent countries of the former Soviet Union. The CSCE and other international institutions can play an important role here, drawing the newly independent countries into a more active dialogue on ways of promoting democratic values and principles. We discussed ways of strengthening the CSCE's role, including the role of the High Commissioner for National Minorities, at the recent Rome ministerial meeting. We must try to find ways of managing and resolving tensions before they erupt into conflict. It will not be easy, but we must try. Potentially explosive cocktails of ethnic and political animosities could spread the fire of intolerance all too easily to other countries and regions. We need to deal with it together in the spirit of partnership. And we need to deal with it urgently. Our security is indivisible and we must ensure it by vigorous and imaginative efforts aimed at promoting peace for all as well as rights for all.

Our Summary:
The international community is starting to look beyond the conflict in the former Yugoslavia, and is becoming increasingly aware of the true magnitude and seriousness of ethnic and political conflicts in the former Soviet Union. The disputes in the Transcaucasus, central Asia, and elsewhere in the former Soviet Union are causing great suffering for those involved. They are a source of legitimate concern to the Russians, who are worried by clashes close to their borders. They are concerned too about the safety of ethnic Russians living in those countries where there are conflicts, and about the economic and social consequences of refugees fleeing to Russia (whose potential number is estimated at more than 2m). They believe that the problems in some of the former Soviet republics are either ignored or misunderstood. The Russians fear that if they do nothing, the conflicts will worsen and probably spread. They argue that they have a legitimate concern in trying to resolve conflicts that threaten their security. Equally, these disputes are of legitimate concern to Britain, its western partners and the whole international community, both for humanitarian reasons and because such flashpoints risk provoking regional instability. If uncontrolled, they pose a variety of threats to European security, including increased traffic in drugs and arms as well as international terrorism. The British government will not underwrite Russian involvement unless it is sure that peacekeeping troops will not turn into occupying forces. When they are willing for Russian troops to play a peacekeeping role, they want there to be some form of international monitoring. It needs to be looked at on a case by case basis. We also need to

put much more emphasis on preventive diplomacy. (** also **)

Copernicus Summary:
The international community is starting to look beyond the conflict in the former Yugoslavia, and is becoming increasingly aware of the true magnitude and seriousness of ethnic and political conflicts in the former Soviet Union. They are concerned too about the safety of ethnic Russians living in those countries where there are conflicts, and about the economic and social consequences of refugees fleeing to Russia (whose potential number is estimated at more than 2m). They ask why, when the international community is engaged in peacemaking and peacekeeping in many of the world's trouble spots, similar efforts should not be undertaken in the hot spots near them. If uncontrolled, they pose a variety of threats to European security, including increased traffic in drugs and arms as well as international terrorism. Crises in Russia's newly independent neighbours may also complicate Russia's relations with the west: unless they are carefully handled, they risk prejudicing the emerging relationship between us. The British government will not underwrite Russian involvement unless it is sure that peacekeeping troops will not turn into occupying forces. They fear that this would compromise their sovereignty. Peace-keeping in the former Soviet Union, as indeed peacekeeping elsewhere, must comply with internationally recognised principles. All this will help enhance Russian peacekeeping practices. Although the Conference for Security and Co-operation in Europe and the UN lack the resources to put large numbers of peacekeepers on the ground themselves, they may be able to suggest measures to enhance transparency and to monitor the peace process.

Comments:
As is clear from both the above summaries, the main deficiency in both of them is the abundance of unbound pronoun references (such as *they*, *we*). It may be noticed that the document itself uses pronouns to a very large extent unlike all the previous three example presented, and this could be the main reason of the failure of both the approaches.