自 然 言 語 処 理　134－6
（1999．11．25）

情報検索タスクに基づく句表現要約の評価

岡 満美子，上田 良寛

E-mail: {mamiko, ueda}@rsl.crl.fujixerox.co.jp

富士ゼロックス(株) 総合研究所

検索結果のふるい分けのための要約として、句表現要約を提案している。その有効性を検証するため、情報検索をタスクとするタスクベース評価を行った。タスクベース評価手法は試行錯誤の段階であるため、実験規模、検索課題、適合度判定方法、判定の正確さの評価方法の点を再検討し、4 段階の適合度判定、場合分け付きふるい分け効率、適合性スコア等を導入した。これにより、現実の検索シーンを反映し、正確性が高く、研究へのフィードバックが可能な評価方法を考案した。この方法を用いて、句表現要約、冒頭 80 文字、重要文ピックアップ、キーワード列挙の各手法を比較した結果、多くの場合、句表現要約が最も有効であることがわかった。「句」を単位とする要約は、ふるい分けに適切な情報量をもち、内容への手がかりを広くカバーするので、「句」は要約に適した単位であるといえる。

Evaluation of Phrase-represented Summarization Method based on Information Retrieval Task

OKA Mamiko, UEDA Yoshihiro

Corporate Research Laboratory, Fuji Xerox Co., Ltd.

To evaluate the performance of the phrase-represented summarization, we carried out a task based evaluation experiment in information retrieval tasks. We have designed tasks similar to the actual WWW retrieval scene, used fine scale to judge the relevance of the documents, and introduced new measure representing the accuracy of sifting. This method can evaluate the performance more accurately than the methods used in the related works. The result of the experiment shows that the phrase-represented summarization can sift documents more accurately than existing summarization methods in most cases. The reason is that phrase summaries have adequate quantity of information for sifting and cover more clue words.

1. はじめに

検索結果のふるい分けを目的とした“At-a-glance”要約の研究を進めている。そのひとつの実現手法として「句表現要約」を提案した(岡他 1999)(上田他 1999)。本稿では、これを既存の手法と比較し、情報検索での有効性を評価する。

自動要約の評価は難しく、評価方法そのものが研究対象ともなっている。これまでに試みられている手法には、理想的な要約との比較(三上他 1998)、要約を利用するタスクの達成率に基づく評価(Jing, et al. 1998)(Mani, et al. 1998)(望月・奥村 1999)、文章としての読みやすさの評価(Wasson 1998)等がある(奥村・難波 1998)。

我々は、情報検索での有効性を示すという目的から、タスクベースの評価手法を選択した。この手法は、SUMMAC(TIPSTER Text Summarization Evaluation Conference)で採用されるなど重要性が認識されつつあるが、実施例は少なく、まだ試行錯誤の段階である。このため、まず評価方法を検討し、考案した方法により実験を行った。

本稿では、まず考案した評価方法を紹介する。次に、評価実験に基づき、句表現要約の有効性を示す。最後に評価方法を検証する。

2. 情報検索タスクに基づく評価方法

2.1. 比較対象

句表現要約とは、重要概念を含む「句」の並びによって概要を表現する要約手法である。ここで「句」という語は、短く、構造が単純な表現を一語で表すものとして用いている。

この手法は、重要文ピックアップによる要約がふるい分けには適していないという考察に基づいて考案したもので、次のような特徴をもつ。

① 短くて単純な表現により、一目で把握できる。
② 重要ポイントのみを抽出することにより、ふるい分けに適した情報量をもつ。
③ 複数の「句」を並べることにより、文書内容への手がかりを広くカバーする。

このような背景から、主要な比較対象は重要文ピックアップである。また、短い表現、手がかりのカバー率の高さといった特徴が共通することから、キーワード列挙も取り上げる。このように、「文」「句」「単語」といった構成単位の違いを比較の中心とする。

実験においては、他の条件を揃えるため、いずれも単語重要度として tf.idf(Salton 1989)を用い、単位の途中で切らないよう配慮して 60～80 文字に要約長を揃える。また、検索クエリは考慮しない。さらに、WWW 検索エンジンでの内容表示として一般的な、冒頭部分のテキストを要約とする手法も比較対象に加える。長さは 80 文字とする。

表 2.1：要約作成方法と要約例

	作成方法	要約例
冒頭80文字	文書の冒頭の80文字を抜き出す	宝島あれこれ　一味違う今年の月餅　今年の「中秋節」は九月十六日。中秋の名月　旧暦の八月十五日にあたるこの日、台湾では祝日となる。日本では「ススキに月見だんご」が……
重要文ピックアップ	単語重要度の文ごとの総和が大きい順に抽出し、出現順に並べる	また、お茶メーカーなどが「茶月餅」と名付けたお茶入りの月餅を開発して、ウーロン茶だけでなく緑茶、ジャスミン茶風味の月餅が売り出されている。台湾の故宮博物院は、七……
句表現要約	プロトタイプシステムで合成された句を、出現順に並べる	...餡のかわりにアイスクリームを中身にしたアイディア月餅も登場した台湾の「得意傳播公司」が制作した同シリーズ... ...唯一故宮博物院の許可を得たものである
キーワード列挙	重要度が大きい順に単語を抽出し、その順に並べる	月餅、餡、故宮、中秋、一味違、中秋節、売り出、同シリーズ、博物院、ROM、台湾、各メーカー、アイスクリーム、23日、8月23日、8月7日、CD-ROM

各要約の作成方法と要約例を表2.1に示す[1]。

2.2. 評価実験の概要

情報検索タスクに基づく評価実験の枠組を図2.1に示す。

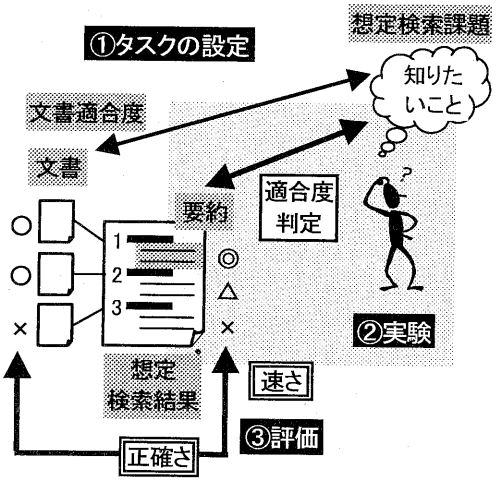

図2.1：情報検索タスクに基づく評価

評価実験は、①タスクの設定、②被験者を用いた実験、③評価、の三ステップからなる。それぞれの概要を以下に示す。

① タスクの設定
「知りたいこと」を具体的に設定する。それについて検索したという想定に基づき、要約を付加した検索結果を用意する。

② 被験者による実験
被験者に、要約を読んで適合度を判定してもらう。

③ 評価
被験者の判定を、正確さと速さのふたつの基準で評価する。

2.3. 新しく考案すべき点

情報検索タスクに基づく評価の従来例(Jing, et al. 1998)(Mani, et al. 1998)は、次の点で検討が必要と思われる。

(1) 実験規模
(2) 検索課題
(3) 被験者による適合度の判定方法
(4) 適合度判定の正確さの評価方法

以下、それぞれの問題点と我々が採用した方法について述べる。

2.3.1. 実験規模

従来例では、検索課題数や対象文書数は多いが、一要約あたりの被験者数は1～2名と少ない。すなわち、個別の分析よりも大量のデータに対する全体的な傾向を重視している。

一方、我々のように一組織内で行う実験の場合には、SUMMACのように大きなコストはかけられない。また、研究の一ステップとしての評価であるため、個々のデータを分析して改良につなげたいという要求がある。

そこで、個々の要約に対して正確なデータをとることを重視して、一要約あたりの被験者を10名とし、課題数は3、課題あたりの文書数は10とした。

2.3.2. 検索課題

従来例は新聞記事を対象としており、我々の「一般ユーザの日常的なWWW検索」という想定とは異なる。また、「知りたいこと」が具体的に設定されておらず、解釈に個人差が出やすい。

そこで我々は、次のような方針をとった。

◇ 実際のWWW文書を用いて、日常的な課題を作成する。
◇ 具体的な検索シーンを設定し、「知りたいこと」をはっきりさせる。

後者は、MLUCEプロトコル(Minel, et al. 1997)の「評価方法の自由度をできるだけ制限する」という制約にも合っている。

作成した課題の概要と、想定結果(全10文書)中での適合文書数を表2.2に示す。ここで「適合文書」とは、「知りたいことの答が書いてある文書」を指す。なお、課題a1とa2は、全く同じ想定結果を利用した。また、実際に被験者に示した課題の例を図2.2に示す。

[1] 要約例の原文書は、http://www.roc-taiwan.or.jp/news/week/1829/108.html

表2.2：検索課題

課題	概要	適合文書数
a1	中秋に月餅を食べる由来	5
a2	月餅の中身の種類	6
b	シンガポール、バンコク間の鉄道の時間、料金、設備	7

課題a2：
中秋直後のある日、同僚のAくんが、中国系の友人に貰ったという月餅を会社に持ってきました。珍しいので、休み時間に切り分けてみんなで食べました。日本では、中秋といえば月見団子ですが、中国などでは月餅を食べるのだそうです。「どうして月餅を食べることになったんだろう？」というのが話題になりましたが、誰も知っている人はいませんでした。そのとき食べた月餅は、馴染みのある甘い月餅のほかに、チャーシューが入ったおかずっぽいものもありました。本場では、月餅の中身は何十種類もあるらしいとBさんが言っていました。私は、月餅について次のことが知りたくなりました。
中華圏で中秋に売り出される月餅の中身には、どのような種類があるのでしょうか？

図2.2：検索課題の例

2.3.3. 被験者による適合度の判定方法

従来例は、適合／不適合の2段階で適合度を判定している。しかし、要約からはどちらとも判断しにくい場合が予想され、無理にどちらかに決めると結果の正確性を損なう。

我々はこのようなケースを考慮し、◎○△×の4段階の判定を採用した。さらに、被験者によるゆれを減らすため、各判定に表2.3のような基準を与えた。

表2.3：適合度の判定基準

判定	基準
◎	要約中に知りたいことの答がある
○	要約中に答に関連する箇所がある
△	要約中に関連する箇所はないが、原文に書いてある可能性は捨てられない
×	要約は知りたいことと関連がない

2.3.4. 適合度判定の正確さの評価方法

従来例は、文書適合度と被験者の判定との一致度を、再現率／適合率で評価している(以下これを「ふるい分け効率」と呼ぶ)。また、総合スコアとしてF-measureを用いている。

ふるい分け効率自体は、情報検索の達成率を表すものとして妥当な尺度である。しかし、前述したように要約からの適合度判定には曖昧性があり、実際のシーンでは、重要度、緊急度といったその時の状況により、どの程度確実な文書を選択するかが変わってくる。そこで、前述した4段階の判定を利用して、状況に応じた場合分けを行うように改良した。

一方、評価結果を問題点の発見につなげるには、ふるい分け効率とは別に、個々の要約の適合度を直接評価できる尺度があった方がよい。そこで、判定がどの程度正しかったかを示す尺度として、適合性スコアを新たに導入した。

場合分け付きふるい分け効率

従来と同様に、再現率／適合率、F-measureを用いる。それぞれの計算式を以下に示す。

$$\text{再現率} = \frac{\text{被験者が適合と判定した文書数}}{\text{実際に適合する文書数}}$$

$$\text{適合率} = \frac{\text{被験者が正しく適合と判定した文書数}}{\text{被験者が適合と判定した文書数}}$$

$$\text{F-measure} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}}$$

ここで「被験者が適合と判定した文書」を、「◎のみ」「◎と○」「◎と○と△」の3種類設定し、「漏れを少なくしたい」「ノイズを少なくしたい」といった状況に応じてふるい分け効率を得られるようにする。

適合性スコア

実際の文書適合度と被験者による判定との一致度をスコアで表す。これを平均することにより、要約手法間の比較ができるほか、文書単位での比較も可能である。スコアの割当てを表2.4に示す。

表2.4: スコアの割り当て

文書適合度	○	○	○	○	×	×	×	×
被験者の判定	◎	○	△	×	×	△	○	◎
スコア	10	8	5	-2	2	-5	-8	-10

3. 実験結果

3.1. 適合度判定の正確さ

場合分け付きふるい分け効率

再現率と適合率の関係を図3.1に、F-measureを図3.2に示す。

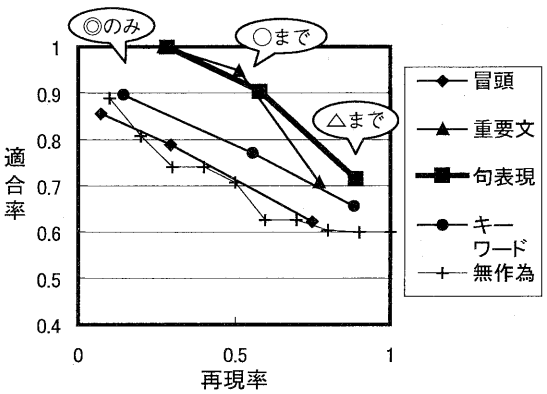

図3.1：再現率と適合率

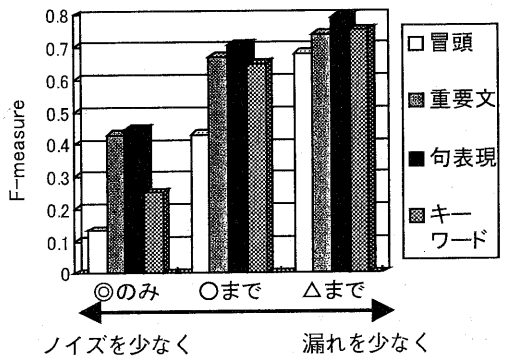

図3.2：F-measure

図3.1、図3.2からわかるように、句表現要約が最もふるい分け効率がよい。また、絞り込んだ場合に漏れが少なく、広く見た場合にノイズが少ないので、バランスのよいふるい分けができる。種々雑多な検索シーンを含むWWW検索に適しているといえる。

適合性スコア

課題別の適合性スコアを図3.3に示す。

図3.3からわかるように、平均では句表現要約が最も高い。課題別では、課題a2および課題bでは句表現要約が高く、課題a1では、全体的に低い中で重要文ピックアップが最も高い。

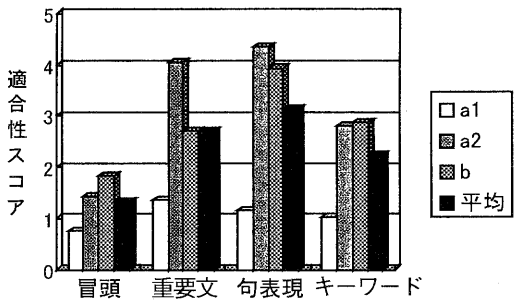

図3.3：適合性スコア

3.2. 適合度判定時間

適合度判定時間を図3.4に示す。なお、課題a1とa2は同時に実験を行ったため、課題aとして両者をあわせた判定時間を示している。

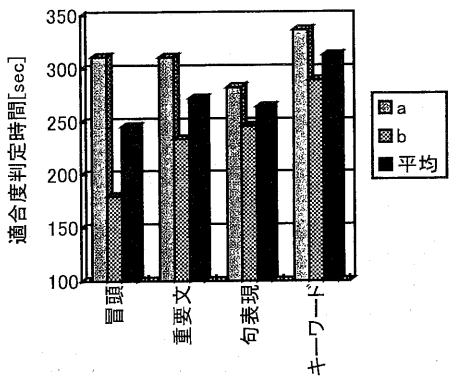

図3.4：ふるい分けの速さ

図3.4からわかるように、課題aでは句表現要約が、課題bでは冒頭80文字が最も短い。課題bは「3つの項目のうち、少なくともひとつが書かれているか」をみるものであり、複数項目を含みやすい句表現要約やキーワード列挙では判定が複雑になるため、時間が長くなったと予想される。このように、課題bは速さを調べるのに適切な課題ではなく、一般には句表現要約が最も速いと予測されるが、複数サンプルでの実証が必要である。

4. 考察

ここでは、「適合文書を正しく判定できる要約」という観点から、構成単位による比較、および課題や文書の影響の考察を行う。

4.1. 構成単位による比較

要約の有効性とその構成単位(文／句／単語)との関連を考察するため、重要文ピックアップ、句表現要約、キーワード列挙の三手法について、以下のような要約の性質を調べた。その結果を表4.1に示す。

◇ 一文／一句／一単語あたりの平均文字数
◇ 要約を構成する文／句／単語数の平均
◇ 適合文書の要約のうち、知りたいことの答や直接的な手がかり語を含むもの(以下、「意図に適合する要約」と呼ぶ)の数
◇ 意図に適合する要約の適合性スコア平均

表4.1：要約の性質

構成単位	文			句			単語		
課題	a1	a2	b	a1	a2	b	a1	a2	b
平均文字数	68.1			18.7			3.7		
平均構成単位数	1.3文			3.9句			17.2語		
意図に適合した要約数	0	3	3	2	6	5	3	5	6
その平均スコア	−	9.43	7.73	7.2	6.83	8	6.77	5.56	6.43

2.1節で句表現要約の特徴として挙げた中で、内容への手がかりのカバー率、ふるい分けに適した情報量の二点について比較する。

内容への手がかりのカバー率

手がかりを広くカバーするということは、要約が任意の検索意図に適合しやすいということである。したがって、手がかりのカバー率は、意図に適合する要約の比率で表すことができる。

表4.1から、意図に適合する要約数は、「句」は「文」に対して2倍以上で、「単語」とほぼ同等である。これより、「句」による要約は「文」よりも広く手がかりをカバーし、これ以上構成単位数が増えてもカバー率はほとんど変わらないことがわかる。

ふるい分けに適した情報量

意図に適合する要約の情報量が適合度を判定するのに適切であれば、そのスコアは高くなるはずである。要約の構成単位と判定しやすさとの関連は、次のように予想される。「単語」は、使われている文脈や単語間の関係がわからないため、判定しにくい。「文」は、全体が課題に関連すればよいが、余分な情報が含まれていると却って判定しにくい。「句」は、必要な情報量を満たしていないと判定しにくいが、余分な情報がないため却って判定しやすい場合もある。

表4.1をみると、平均スコアは「単語」による要約が最も低く、「文」よる要約と「句」による要約とは課題によって異なる。これは上述した予想を裏づける結果になっている。

以上をまとめると、表4.2に示すように、「句」による要約は、手がかりのカバー率と適切な情報量の両方を満たしており、「句」は要約に適した単位であるといえる。

表4.2：単位による比較

	文	句	単語
手がかりのカバー率	×	○	○
情報量	○	○	×

4.2. 課題や文書の影響

3.1節で示したように、適合度判定の正確さ

は全体としては句表現要約がよいが、課題によってばらつきがあった。ここでは、課題や文書の性質と要約手法との関連を考える。

課題の影響

表4.3に示すように、知りたいことの表現レベルにより、適切な要約手法は異なってくる。

表4.3：表現レベルと適切な要約手法

知りたいことの表現レベル	向いている要約手法
短文レベル	句表現要約
重文／複文レベル	重要文ピックアップ
複雑な文間関係	どの手法でも困難

ただし、答を予想しやすい課題、手がかり語を思いつきやすい課題では、答の断片でも判定が可能なため影響は小さいと思われる。

文書の影響

知りたいことに関連する部分が要約として抽出されるかどうかは、それが主要な話題であるかどうかに依存する。重要文ピックアップでは一文程度しか抽出されないため(表4.1参照)、話題が複数ある場合やサブトピックの場合には、句表現要約やキーワード列挙の方が抽出されやすい。しかし、かなり周辺的な話題になるとどの手法でも困難である。

以上の観点から、次のようにまとめられる。

■ 重要文ピックアップが向いている場合
知りたいことが重文／複文レベルで表される比較的難しい課題。文書の中心的な話題である場合に限られる。

■ 句表現要約が向いている場合
知りたいことが短文レベルで表現し得る課題。文書中では多少周辺的な話題であってもよい。

■ どの手法でも困難な場合
知りたいことが複雑な文間関係で表現され、かつ難しい課題。例えば、原因、方法などを問うもの。また、文書中でかなり周辺的な話題である場合。

今回の課題をこれらの観点からみると、a1はどの手法でも困難な課題、a2、bは句表現要約が向いている課題であった。中心的な話題ではない文書はどの課題にも含まれ得るので、重要文ピックアップが向く場合は非常に限られる。したがって、どの手法でも困難な場合を除けば、一般に句表現要約が最も有効であるといえる。

5. 評価方法の検証

本研究と従来の情報検索タスクに基づく評価例との比較を表5.1に示す。

我々の方法の利点と問題点を以下にまとめる。

表5.1：評価方法の比較

項目		本研究	[Jing他 98]	SUMMAC 98	[望月他 99]
規模	検索課題	3	4	20	10
	課題あたりの文書数	10	10	50	20
	被験者	社員40名	学生12名	情報アナリスト21名	情報系学生21名
	要約あたりの被験者数	10名	1名	1～2名	1～3名
設定	対象	WWW	新聞記事	新聞記事	新聞記事
	検索課題の設定	具体的な検索シーン	トピック	トピック	トピック
	課題と想定検索結果	新たに課題を作成し、WWW検索結果から選択	TRECテストコレクションより選択	TRECテストコレクションより選択	BMIR-J2より選択
実験	適合度判定	◎○△×の4段階	適合／不適合	適合／不適合	適合／不適合 全文の参照可
	評価尺度	正確さ、速さ	正確さ、速さ	正確さ、速さ	正確さ、速さ 読みやすさ
評価	正確さの尺度	場合分け付きふるい分け効率、適合性スコア	ふるい分け効率	ふるい分け効率	ふるい分け効率

5.1. 利点

■ 各要約の適合度判定の正確さ

要約あたりの平均の判定結果と、最初の1〜2名の判定結果との間には、ずれがあった。被験者を 10 名ずつとすることにより、正確性の高い実験ができたといえる。

■ 課題解釈のゆれの少なさ

被験者に原文書の適合度を判定してもらったところ、想定との一致度は 93%であり、SUMMAC 98 の 69%と比較して格段に高くなった。これは、検索シーンを具体的に設定したことによると思われる。

■ 曖昧性を考慮した適合度判定

適合度判定を4段階にし、ふるい分け効率に場合分けを導入した。これにより、広く見たい、絞り込みたいといった状況に応じて要約手法の性質を表現できた。

■ 適合性スコアの導入

適合性スコアによって文書ごとの比較ができ、要約手法の特徴や問題点の分析といった、研究へのフィードバックが可能になった。

5.2. 問題点

- ◆ 各被験者は二手法ずつしか評価していないため、個人差の程度を全体として比較することができなかった。
- ◆ 検索課題、文書の選択にあたり、4.2 節で考察したような特徴を考慮すれば、より効果的な実験ができた。

6. まとめ

本研究では、情報検索タスクに基づく評価において、現実の検索シーンを反映し、正確性が高く、研究へのフィードバックが可能な評価方法を考案した。また、この方法を用いて、句表現要約が既存の手法よりも検索結果のふるい分けに適していることを示した。

現在、評価結果の分析により明らかになった問題点に基づいて、句表現要約の品質向上を行っている。今後は、今回不十分だった速さの再評価と、要約の読みやすさや読む際に感じるストレスの評価を行う予定である。

参考文献

上田, 岡, 小山, 宮内 (1999). "句表現要約に基づく要約システム." *言語処理学会第5回大会発表論文集*, pp361-364.

岡, 小山, 上田 (1999). "句表現要約の句合成手法." *情報処理学会自然言語処理研究会報告 99-NL-129*, pp101-108.

三上, 山崎, 増山, 中川 (1998). "文中の重要部抽出と言い換えを併用した聴覚障害者用字幕生成のためのニュース文要約." *言語処理学会第4回年次大会併設ワークショップ「テキスト要約の現状と将来」論文集*, pp14-21.

望月, 奥村 (1999). "情報検索タスクに基づいた評価による要約手法の比較." *情報処理学会自然言語処理研究会報告 99-NL-132*, pp41-48.

Jing, H., Barzilay, R., McKeown, K. and Elhadad, M. (1998) "Summarization Evaluation Methods: Experiments and Analysis." In *Intelligent Text Summarization*. pp51-59. AAAI Press.

Mani, I., House, D., Klein, G., Hirschman, L., Obrst, L., Firmin, T., Chizanowski, M., and Sundheim, B. (1998) "*The TIPSTER SUMMAC Text Summarization Evaluation.*" Technical Report MTR 98W0000138, MITRE Technical Report.

Minel, J.-L., Nugier, S. and Piat, G. (1997). "How to Appreciate the Quality of Automatic Text Summarization? Examples of FAN and MLUCE Protocols and their Results on SERAPHIN" In *Proc. of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pp. 25-30.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Publishing Company, Inc.

Wasson, M. (1998). "Using Leading Text for News Summaries: Evaluation Results and Implications for Commercial Summarization Applications." In *Proc. of COLING-ACL'98*, pp1364-1368.