Text compaction for display on very small screens

Simon CORSTON-OLIVER
Microsoft Research
One Microsoft Way
Redmond WA 98052, USA
simonco@microsoft.com

Abstract

Very small screens, such as the LCD displays on cellular telephones, pose unique problems for the display of textual messages. This paper presents a collection of techniques used to compact email text for display on mobile devices. These techniques range from simple string manipulations to more sophisticated linguistic processing. The techniques have been implemented in a commercial product for four languages: English, French, German and Spanish.

1. Introduction

We are in the midst of a resurgence of interest in text summarization. Although some work is being done on summarizing non-linguistic sources (see for example McKeown et al 1995), most current research focuses on the summarization of a single document or a collection of documents. The summaries produced are typically classified as either *indicative*, i.e. providing a good indication of the content of a source document, or *informative*, i.e. providing the essential information of the source document so that reading the summary is a practical substitute for reading the original document. Most researchers use an assortment of statistical techniques and discourse analysis to determine which portions of a document to extract (see Sparck Jones 1998 for an overview of recent work on summarization). These extracts are then presented to the user, perhaps after some massaging to improve coherence. It is typical to set an upper limit on the number of words in a summary, either as an absolute limit or as a percentage of the number of words in the source text.

Several recent studies move beyond the extraction of sentences or other relatively large textual units. Banko et al. (2000) and Witbrock and Mittal (1999), for example, produce document summaries in the style of newspaper headlines. These summaries consist of a single sentence or even less than a sentence by selecting words from across an entire document. Jing and McKeown (2000) and Knight and Marcu (2000) describe techniques for producing extended summaries that encompass textual units of varying sizes. Both studies combine and reduce content drawn from multiple textual units into a single sentence.

The closest analog to the work presented in this paper is the text reduction performed by Grefenstette (1998) who produces telegraphic representations of every sentence in a document by deleting elements of a sentence determined to be relatively unimportant on the basis of a shallow syntactic analysis. These telegraphic representations are motivated by user interface considerations: text reduction enables a visually impaired reader to skim a document being read aloud.

Text summarization and text reduction are inherently lossy processes since both necessarily involve decisions about what elements of a document can safely be omitted. In most scenarios in which summarization or reduction is employed, users have the option of reviewing the original document. Some loss of content is thus acceptable, since the user can always review the source text to find the full information. In the system that Grefenstette (1998) describes, for example, the user can turn a knob to vary the rate of compression and back up to repeat a stretch of a document with less reduction or even no reduction. Similarly, summaries in the style of a newspaper headline are indicative of the content of documents that readers might wish to review in their entirety.

In this paper, I describe a component of Microsoft Outlook Mobile Manager[1], a commercial product that performs text compaction on selected

[1] Available for download from http://microsoft.com.

email messages which it then routes to a mobile device, such as a cellular telephone or a pager. Two factors constrain the presentation of email messages on a small LCD screen. On the one hand, since most mobile devices are capable of receiving electronic messages but not of responding to an email server, the recipient of the message must be able to understand it by reading only the compacted form. Even if it were possible for the user to request the full text of the message from the email server, it would be impractical to read any but the briefest of uncompacted messages on the extremely small display[2]. We must therefore minimize the loss of content. On the other hand, we need to reduce the form of the message as much as possible so that it will fit on the small display.

Since the strategies described here are intended to keep the loss of content to an absolute minimum, I use the term *compaction*, as opposed to *reduction* or *summarization*, both of which imply deciding what content can be omitted. The focus here is on producing a more compact form of a message. Following a brief discussion of the overall architecture of Microsoft Outlook Mobile Manager, I proceed to a discussion of the set of techniques used to reduce the text.

2. Microsoft Outlook Mobile Manager

Microsoft Outlook Mobile Manager monitors a user's incoming email[3]. On the basis of an initial session in which the user trains a classifier, certain messages are selected to be routed to the user's device. (The classifier is similar to the one described in Sahami et al 1998). The email message is parsed to remove formatting information such as ">" symbols indicating forwarded mail and to extract essential information such as the name of the sender, the title of the message, and the date and time sent. The title and body of the message are then analyzed using the Microsoft natural language processing system (Heidorn 2000). During this phase, a morphological and syntactic analysis is performed. A postprocess then proposes a text compaction for each leaf node in the syntax tree. Each proposal consists of a descriptor accompanied by the text of the leaf node at each of three levels of compaction. Linguistic analysis is performed for English, French, German and Spanish. Some of the compactions proposed apply across all languages, others, such as the handling of morphological inflection, are language-specific.

The message router takes into consideration the user's specified preference for the degree of compaction (ranging from none to the most extreme compaction possible) and the number of characters available on the user's mobile device, and weighs the benefit of each compaction proposed by the natural language processing system. Some proposed compactions are accepted, others are deemed too severe given the user's specified preferences or unnecessary because the message will fit in the limited space without the compaction.

Finally, the message is transmitted to the device in chunks, with each chunk filling the available space on the small screen. For the remainder of this paper I focus on the text compaction strategies.

3. Text compaction

We process the text of an email message on a sentence-by-sentence basis. A simple rule inspects the syntactic parse for each sentence. If the parse appears to contain a relatively high proportion of punctuation characters (source code, tabular statistical data etc), then we flag it as not being suitable for text compaction and proceed to the next sentence.

For sentences that survive this initial inspection, we recommend various text compactions. For each leaf node in the syntax tree, we perform additional analysis based on the part of speech of the leaf node. We return a descriptor, referred to as the SHORTTYPE, and three levels of compaction: the LONGFORM, the COMPRESSEDFORM, and the CASENORMALIZEDFORM. It should be noted that leaf nodes in the linguistic analysis performed by the Microsoft grammars can contain more than one word in the case of lexicalized collocations and minor phrase types, such as numbers spelled out in full or personal names.

The LONGFORM contains the text of the original string. The only compaction performed to produce the LONGFORM is to reduce multiple leading spaces to a single leading space. The COMPRESSEDFORM contains a compacted representation of the original

[2] Many newer mobile devices have larger screens, but it will be some years before these new devices supplant current telephones and pagers.

[3] The design and implementation of the architecture described in this section was performed by Sharad Mathur.

string. The type of compaction is noted in the SHORTTYPE descriptor. The CASENORMALIZEDFORM contains a modified version of the COMPRESSEDFORM with certain characters removed or normalized. I describe the process of removing characters and normalizing case in sections 3.1 and 3.2 below, and then proceed to a description of each linguistic compaction strategy.

3.1 Character removal

The default compaction strategy involves the removal of certain characters. In the discussion below, I use the term "vowel nucleus" to refer to a single vowel or several contiguous vowels. Note that this correlates only weakly to the notion of a syllable nucleus in phonology. The character removal strategies are different for each language.

English. All word-internal vowels are deleted. For example, *example* becomes *exmple*.

French. Even-numbered word-internal vowel nuclei consisting of one or two vowels are deleted. Longer nuclei are retained because they frequently contain a syllable boundary. The derivational suffix *–ment* is compacted to *–mt*. For example, *sérieusement* becomes *sérieusmt* 'seriously'. Note that since the sequence of three vowels *ieu* is invisible to vowel deletion, the following *e* is deleted. The sequence *en* is deleted from the derivational suffix *–ment*.

German. The consonant cluster *ck* is simplified to *k*, *sch* is simplified to *sh* except in the word-final sequence *–schen* (word-final *–s* followed by the diminutive suffix *–chen*), and every second word-internal vowel nucleus consisting of one or two vowels is deleted, with one exception: vowels that follow the letter *s* and precede the consonant cluster *ch* are not deleted because doing so would result in the sequence *sch*, a common syllable-initial consonant cluster. After the deletion of every second medial vowel, the letter *e* is deleted if it precedes a word-final *l*, *m*, *n*, or *r* and follows a consonant. In remaining instances of the sequence *Consonant + ie + Consonant*, the sequence *ie* is simplified to *i*; Finally, the letter *u* is deleted in the word-final string *Consonant + ung* and the sequence *–ein* is replaced with the homophonous digit *1* except in certain words such as *Kaffein* 'caffeine' and *Codein* 'codeine' etc.). By way of example, consider the verb *vorbeieilen* → *vorbeieiln* 'to rush by'. Note that the sequence of four vowels *eiei* is retained and the *e* that follows a consonant and precedes the word-final *n* is deleted. The word *Kartoffelstampfer* 'potato masher' is reduced to *Kartffelstmpfr* by deleting every second vowel nucleus and by deleting the letter *e* before the word-final *r*.

Spanish. The Spanish vowel deletion strategy is sensitive to word length: words with only two word-internal vowel nuclei have the second vowel nucleus deleted, for example *muchos* → *muchs* 'many/much'; longer words often have derivational morphology preceding the stem and inflectional morphology at the end of the word. We retain the first two word-internal vowel nuclei, which frequently results in the first one or two vowels of the stem being retained, but delete all subsequent vowel nuclei, for example *desarrollando* → *desarrllndo* 'developing'.

The character deletion strategies for each language were refined through a process of informal feedback by native and near-native speakers of each language. Native speakers of French, German and Spanish tended to agree that the more aggressive deletion strategy tolerated by English speakers was not in keeping with orthographic practices for the other languages. Several non-native speakers of English commented on vanity license plates on cars and signage conventions as orthographic precedents for vowel deletion in English.

It should be noted that the problem of deleting characters could be viewed from the perspective of information theory. Certain characters are predictable given preceding context and can therefore be omitted. Although a more mathematical strategy informed by information theory might yield higher rates of compaction, it would not be transparent to naïve users. We have tried as much as possible to delete characters in ways that allow users to recover them easily.

3.2 Case Normalization

The CASENORMALIZEDFORM is based on the COMPRESSEDFORM. Spaces are deleted and word-initial letters are capitalized to aid the reader in

locating word boundaries. Acronyms are left as all capitals; in all other words letters that occur word-medially are converted to lower case. Occasionally, a leading space must be preserved to aid in comprehension, for example to separate two adjacent numbers.

3.3 Linguistic text compaction strategies

If the text corresponding to a leaf node occurs in a list of arbitrary substitutions, then the substitution is used as the value of the COMPRESSEDFORM. Characters in arbitrary substitutions are not deleted, since one function of the substitution mechanism is to override character deletions that might lead to embarrassing or undesirable output. One example of arbitrary substitution is the use of the ampersand for English *and*, German *und* and French *et*. (The Spanish *y/e* is already one character and so does not benefit from compaction). Similarly, German *deutsche Mark* → DM, English *inch*→" and so on.

If no arbitrary substitutions apply, we check to see whether any compactions can be performed. Only compactions applicable to the part of speech of the leaf node are examined. In the following sections I briefly describe the compaction strategies used for each part of speech. For the sake of brevity, examples will be restricted to English except when the strategy applies only to another language.

3.3.1 Noun

There are more compaction strategies for nouns than for any other part of speech. We distinguish company names, days, months, relative dates, absolute dates, email addresses, URLs, personal proper nouns, numbers, telephone numbers, pronouns, money, geographical place names, other proper nouns and default. In the discussion below we present the strategy used to set the COMPRESSEDFORM. Unless otherwise noted, the LONGFORM is set to the original text of the node and the CASENORMALIZEDFORM is derived automatically from the COMPRESSEDFORM.

Company names

The company type can be deleted if present. All punctuation separating elements of the company name can also be deleted.

Examples:
Microsoft, Corporation → *Microsoft*
IBM Ltd. → *IBM*

Days

Days of the week can be compressed to a two or three letter form.

Examples:
Monday → *Mon*
Tuesday → *Tue*

Months

Months can be compressed to a two or three letter abbreviation.

Examples:
November → *Nov*
October → *Oct*

Relative dates

For English, we reused existing code that calculates dates whose value is given relative to the current temporal reference point[4]. For email, the temporal reference point is generally the date that the message was sent. The order of elements in dates and the particular separator used (a forward slash or a period) is determined by language.

Examples (assuming a message is sent on Monday 15 January 2001):
next Monday → *1/22/2001*
the day after tomorrow→*1/17/2001*

Absolute dates

Given a month (as a number or as a word) and year with no day of month, we produce a numeric representation of the month and the year.

Examples:
November 2000 → *11/2000*
Jan 2000 → *1/2000*
2/2001 → *2/2001*

Given a day of the week and full date, we produce a numeric date, with the day of the week omitted.

Examples:
Monday 15 January 2001→*1/15/2001*
Tue 1/16/2001 → *1/16/2001*

[4] This feature is not currently implemented for languages other than English.

Note that with all dates the year can be omitted if it is the same as the current year.

Email addresses and URLs
No compression of any kind is performed. Some mobile devices have web browsing capabilities, so it is conceivable that a user might want to follow a URL embedded in an email message.

Personal proper nouns
If we have a given name or family name, all titles can be deleted. If we have a family name, a given name can be deleted. Examples:

Dr Jane Smith → Smith
Jane Smith → Smith

The exception to this is when we have coordinated titles or first names. For example:

John and Mary Smith → John & Mary Smith
Mr and Mrs Smith → Mr & Mrs Smith

Since proper names are often uncommon lexical items, we do not delete vowels in family names. If a family name has several components, we select only the last one, as the following Spanish example illustrates:

José Gil de Castro → de Castro

Numbers
Numbers spelled out in full can be replaced with digits. In English, multiples of one thousand, one million and one billion are expressed using K, M, and B respectively.

Examples:
twenty-one books → 21 books
nine thousand dollars → $9K[5]

Ordinal numbers in French are indicated by the suffix "e", e.g. *deuxième → 2e* 'second' and in German by the addition of a period e.g. *dritte*→3. 'third'.

[5] Note the interaction with the Money compaction strategy: the word "dollars" is replaced by "$" and the output is reordered.

Telephone numbers
Parentheses, hyphens and other punctuation in telephone numbers can be deleted.

Examples:
(425) 703-7371 → 4257037371
425-703-7371 → 4257037371

Pronouns
In English, the pronoun *you* can be replaced with the letter U. Spanish *Usted* and *Ustedes* can be replaced with *Ud* and *Uds* respectively. German pronouns that end in homophones of the numeral one or its inflected forms can replace *ein* with *1*, e.g. *meinem → m1m* 'my'.

Money
Currency words can be replaced with short forms, e.g. *franc*→ Fr. Certain currency symbols must be reordered, e.g.

nine thousand dollars → $9K
ten francs → 10Fr

Geographical place names
States and countries are replaced with their conventional abbreviations. All other Geographical place names receive the default compaction.

Examples:
South Dakota, United States of America → SD, USA
New Zealand → NZD

Other proper nouns
Proper nouns that are not personal names or geographical place names are likely to be low-frequency items such as the titles of books or songs. We therefore do not apply character deletions aside from space-removal.

Default
The default strategy is to set COMPRESSEDFORM equal to LONGFORM, and then to allow character deletion and case normalization to apply.

3.3.2 Punctuation
For leaf nodes that consist solely of punctuation symbols we distinguish essential punctuation such as sentence-final question marks or periods from

non-essential punctuation such as commas separating list elements, or leading bullets.

3.3.3 Determiner
In English, all definite and indefinite determiners are deleted, e.g.

I bought a book → I bought book
I saw the movie → I saw movie

In French, German and Spanish, definite determiners are deleted. Indefinite determiners in these languages, however, are homophonous with the number one, and therefore cannot be deleted. They are instead replaced with the numeral *1*. The German negated indefinite article *kein* and its inflected forms is replaced with *k1*. Examples:

French: *un livre → 1 livre* 'a/one book'
German: *kein Mann → k1 Mann* 'no man'

There is one special case in French in which the definite article must not be deleted: when the distinction between the comparative and the superlative would be lost. For example

C'est plus vite 'it's faster'
C'est le plus vite 'it's the fastest'

Note also that in many European languages, definite articles are etymologically related to pronouns. We do not delete pronouns either in independent or in clitic positions. The syntactic analysis allows us to distinguish the two environments in which these pronouns occur. The text book example of the French ditransitive *donner* 'to give' illustrates this nicely. The clitic pronoun 'le', which is homophonous with the definite article 'le', is not deleted.

Je le lui ai donné 'I gave it to him'.

3.3.4 Preposition
The English preposition *through* can be replaced by *thru*, and *to* by *2*, *for* by *4*. For the sake of clarity, the abbreviations *2* and *4* are not used before or after numerals. All other prepositions in English and the other languages undergo default compaction.

3.3.5 Verb
Two standard substitutions are made in English: *be→ B*, and *are→ R*. For all other verbs only default compaction applies.

3.3.6 Adverb, Adjective, Conjunction, Interjections
If there are no relevant word substitutions (e.g. *and→ &*), then the default strategy is to set COMPRESSEDFORM equal to LONGFORM, and then to allow character deletion and case normalization to apply.

4. Worked examples

The precise output sent to the mobile device is influenced by the level of compaction that the user has selected and the amount of space available on the screen of the device. For the sake of exposition, I present maximally compressed examples here, i.e. examples in which all possible compactions have been selected, including the deletion of spaces and other characters. Approximate translations into English are given for the French, Spanish and German examples.

French

Source text

Pour autant, on ne peut manquer de s'interroger sur le choix qui préside à l'élaboration de ce classement, puisque les personnes interrogées ont à faire leur choix, selon la méthode des quotas, dans une série de cinquante noms arbitrairement proposés.

Compacted output

PourAutantOnNePeutManqrDeS'InterrgerSurChoixQuiPrésdeÀÉlabratnDeCeClassmentPuisquePersnnesInterrgéesOntÀFaireLeurChoixSelnMéthdeDesQuotsDs1SérieDe50NomsArbitrremtPropsés.

English translation

Moreover, one can't help wondering about the choice that determines the elaboration of the classification since the people interviewed have to make their choice following the quota method in a series of 50 arbitrarily proposed names.

This example, taken from a Usenet discussion illustrates a number of text compaction strategies for French. Vowels have been deleted, spaces removed and the initial letter of each word has been

capitalized. (Note that accents are retained in capital letters to improve readability.) The definite articles have been deleted, e.g. the *le* preceding *choix* and the *l'* before *élaboration*. The indefinite article *une* before *série* has been replaced with the numeral *1*, and the number *cinquante* has been replaced with the digits *50*. The abbreviation *Ds* is an arbitrary substitution for the preposition *dans* 'in'. The text has been reduced from 251 characters to 170 characters, a reduction of approximately 32%.

German

Source text

Hunderttausend Geschäften und Haushalten wurde am Mittwoch erstmals gezielt der Strom abgedreht.

Compacted output

100000Geshftn&HaushltnWurdeAmMiErstmals GezltStromAbgedrht.

English translation

For 100,000 businesses and households the power was specifically switched off for the first time on Wednesday.

The compaction of this German extract from an online news source illustrates the conversion of a spelled out number, *Hunderttausend*, to a numeric form, 100000. The consonant cluster *sch* is compacted to *sh*, the conjunction *und* is replaced with the ampersand, definite articles are removed, the day *Mittwoch* 'Wednesday' is replaced with the two letter abbreviation *Mi*, and so on.

English

Source text

The problem of automatic summarization poses a variety of tough challenges in both NL understanding and generation.

Compacted output

PrblmOfAutmtcSmmrztnPssVrtyOfTghChllngsIn BthNLUndrstndng&Gnrtn.

The English compaction of this excerpt from an email message containing the call for papers for an NAACL'01 workshop clearly shows the effect of deleting all medial vowels. In addition, the conjunction *and* has been replaced by the ampersand, and definite and indefinite articles have been deleted. The text has been compacted from 115 characters to 63 characters, a reduction of approximately 45%.

Spanish

Source text

Este viernes 18 de junio se está desarrollando en Madrid lo que se ha venido a denominar "Reclama las Calles".

Compacted output

Este18.6SeEstáDesarrllndoEnMadrdLoQueSeH aVendoADenomnrReclamaLasCalles.

English translation

This Friday, the 18th of June, what has come to be called "Take Back the Streets" has been developing in Madrid.

This example, from a web page, illustrates date compaction. The day of the week *viernes* 'Friday' has been deleted and the month and day of month rendered in the language-appropriate format as 18.6. A conservative strategy has been followed for the title "Reclama las Calles" 'Take Back the Streets'. Although this is clearly a proper noun phrase, no compressions could be found. We therefore erred on the side of caution and did not delete any vowels. The text has been reduced from 110 characters to 71 characters, a reduction of approximately 35%. Four further characters could be omitted by eliminating the sentence-initial *este* 'this'.

5. Evaluation

Since the primary motivation for the text compactions was to satisfy display constraints, we decided to measure the rate of compaction and to verify that users could decipher the original message from the compacted text.

We focused on English text with the maximal text compaction. Five evaluators, who had not previously been exposed to the Outlook Mobile Manager output, were each given the same set of 100 compacted sentences taken from personal email. The order of presentation was randomized for each evaluator. The topics of the email messages ranged from planning social events to corporate memos.

The sentences in the sample contained an average of 97.7 characters, with a standard deviation

of 42.7 characters. The compacted sentences contained an average of 57.1 characters with a standard deviation of 25.6 characters. Per-sentence compression ranged from 26.3% to 54.4%. On average, the sentences were compressed by 41.6% with a standard deviation of 4.6%. The following sentences illustrate the compacted sentences that the evaluators attempted to decipher.

IKnwThtICnCrteTSCFrmMyWin2kServerMchneBtIs ThtGng2BNwstVrsnThn(SnceMyWin2kServerSystmI sUp-To-Dte)OrDoINdSprteUpdte4TSC?

DrctDepositPymntsWllBAvlbleInYrAccntWthn3Bsns sDysFrmPymntDteBlw.

ThghWeHdExpctd2BAble2SndThsItm2UWe'veSnceF ndThtItIsNtAvlbleFrmAnyOfOurSrcs@ThsTme.

We measured the edit distance between the actual original form and the evaluators' best guesses as to the original text, counting additions, deletions and substitutions of words. Variations in case and punctuation were ignored.

It must be emphasized that the task confronting the evaluators was more difficult than the intended use of the compaction techniques. In normal use readers would encounter acronyms and technical terms from their own domain of specialty. For example, the string *WhenIRASdIn* is difficult to decipher unless the reader is familiar with the acronym "RAS", a dial up network connection used to connect to an ISP. Also, in normal use, the title of the email message would aid in disambiguation. Finally, in normal use the text would be compacted just enough to fit within the display of the mobile device, whereas for this evaluation all words were subjected to the maximum compaction.

Despite these caveats, the evaluators were very successful in decoding the original messages. Table 1 gives a breakdown of the errors per evaluator. Important errors within each category are labeled "Salient". Salient substitutions, deletions and insertions are those other than typos, alternative spellings and articles.

The most frequent error was a failure to hypothesize a deleted article. For example one evaluator, presented with a string beginning *NpeVrsnIGt*, hypothesized "Nope, version I got", omitting the definite article that occurred in the original text. Other deletions were uncommon.

Insertions of words other than articles were rare. For example, for a sentence beginning *Thnks4Shppng@Amazon.Com*, one hypothesis began with the words "Thank you", i.e. a substitution of "thanks" for "thank" and an insertion of the pronoun "you".

Substitutions were the most common form of salient error. Morphological alternations were counted as salient substitutions, e.g. "get" for "got", "come" for "came". Other salient substitutions involved important content words e.g. "caves with sea otters" for "coves with sea otters", "the same Web browser compatibility problems" for "some Web browser compatibility problems". Salient substitutions, such as "caves" for "coves" that do not materially affect the comprehension of the sentence are nonetheless included in the salient error counts below. The non-salient substitutions included typos, e.g. "nuisa[n]ce", alternative spellings, e.g. "altho[ugh]" and "Fri[day]", and articles, e.g. "a" for "the".

There were precisely 1,800 words in the original sentences. As Table 2 shows, the rate of salient errors ranged from 1.1% to 2.2% across evaluators.

Table 1 Edit distance metrics

Evaluator	Deletions		Substitutions				Insertions	
	Articles	Salient	Typos	Alternative spelling	Articles	Salient	Articles	Salient
1	43	9	6	10	7	28	1	2
2	45	4	2	11	4	23	3	5
3	28	5	1	14	7	30	2	5
4	18	3	2	12	5	14	9	9
5	15	1	2	6	7	18	8	1

Table 2 Salient error rate

Evaluator	Salient errors	Error rate (salient errors/1800)*100
1	39	2.2%
2	32	1.8%
3	40	2.2%
4	26	1.4%
5	20	1.1%

Table 3 Sentence error rate

Evaluator	Exact match (all errors)	Exact match (salient errors only)
1	42	74
2	51	81
3	50	71
4	50	82
5	59	85

Table 3 gives the number of sentences correctly deciphered per evaluator. Excluding non-salient errors, there are 50 sentences for which all evaluators deciphered an exact match, and only four sentences for which no evaluator did.

6. Future directions

The current compaction strategies focus on text compactions for the leaf nodes of a syntax tree. Much of the information that is used to suggest text compactions is a side effect of the morphological and syntactic analysis. For example, morphological analysis returns the numeric value for a number spelled out in full, allowing us to render *fifty-two* as *52*. Similarly, a side-effect of syntactic analysis is the disambiguation of the part of speech of the token *le* in French that allows us to determine whether it is a definite article that can be deleted or a clitic pronoun that should be retained. A set of rules that operate after morphology and before the major syntax rules identifies the internal structure of proper names and other minor phrase types. These rules heuristically identify elements as family names or given names that might not be specified as such in the lexicon, giving us the ability to discard given names and titles while retaining family names.

Of course, many of the text compaction strategies presented here could be approximated by less expensive techniques than performing a full syntactic analysis. For the current implementation, we have effectively treated the existing broad-coverage grammars as a convenient library of analysis routines. In future work we intend to explore ways to extract more value from the full syntactic analysis. It may be possible to apply techniques that rely on an analysis of long distance dependencies. Consider the following example.

There is a systematic ambiguity in English in sentences that contain a speech act verb in the main clause and a complement clause. If the complement clause contains a pronoun that matches the subject of the main clause in gender, person and number, then two readings are possible. In one reading the subject of the complement clause is coreferential with the subject of the main clause. In the other reading, the subject of the complement clause is not coreferential, as the following example illustrates:

$John_i$ said that $he_{i,j}$ was not feeling well.

Compare this to the following sentence, in which the mismatch between the gender of the subject of the main clause and the gender of the pronominal subject of the subordinate clause precludes a coreferential interpretation:

$John_i$ said that she_j was not feeling well.

Given a syntactic analysis, and without even performing full anaphora resolution, we could identify this possible long distance relationship and propose a text compaction in which the pronominal subject of the complement clause was omitted, yielding the following text compaction:

John said that was not feeling well.

The complementizer *that* can be deleted, as can the verb *was* which merely encodes default tense-sequencing information between the main and complement clauses, yielding

John said not feeling well.

It may be possible to extract more value from the morphological analysis performed as part of the syntactic parsing. For example, perhaps vowels in inseparable prefixes in German should be retained, or perhaps at least the first vowel of every stem in a German noun compound should be retained

Finally, we can make inferences about the identifiability of discourse referents by examining a user's previous email history. For example, a proper name could be compressed if it was the name of a

person with whom the user has frequent email contact, but a proper name that appears to refer to an unfamiliar person could be left uncompressed.

7. Conclusion

A collection of text compaction strategies serves to reduce the amount of space required to display an email message, enabling a user to view it on the small screen of a mobile device such as a cellular telephone or a pager. Because the user can view the summary but cannot revert to the original document forces, we must compact text while keeping the loss of content to a minimum. Finally, users are very successful at deciphering the compacted text.

Acknowledgements

My thanks go to Sharad Mathur for helpful discussion to clarify the specification of the text compactions. Karin Berghöfer, Margaret Salome, Roger Billerey-Mosier, all of the Butler Hill Group, and the members of the Microsoft Outlook Mobile Manager team provided invaluable testing feedback and linguistic intuitions. The code to calculate relative dates was written by Patti Schmidt. Monica Corston-Oliver of the Butler Hill Group and five anonymous evaluators helped with evaluation. Finally, thanks to the members of the MSR NLP group for fine-tuning grammars and helping with translations.

References

Banko, Michele, Vibhu Mittal and Michael Witbrock. 2000. Headline Generation Based on Statistical Translation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong. 318-325

Grefenstette, G. 1998. Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. In *Intelligent Text Summarization, AAAI 1998 Spring Symposium Series*, Stanford California. 111-117.

Heidorn, G.E., 2000. Intelligent writing assistance. In R. Dale, H. Moisl and H. Somers (Eds.). *Handbook of Natural Language Processing*. New York, NY. Marcel Dekker. 181-207.

Jing, Hongyan and Kathleen McKeown. 2000. Cut and Paste Based Text Summarization. *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2000)*. 178-185.

Knight, Kevin and Daniel Marcu. 2000. Statistics-Based Summarization—Step One: Sentence Compression. *The 17th National Conference on Artificial Intelligence (AAAI--2000)*. 703-710.

McKeown, K., J Robin and K. Kukich. 1995. Generating concise natural language summaries. *Information Processing and Management* 31:703-733.

Sahami, M., S. Dumais, D. Heckerman and E. Horvitz. 1998. A Bayesian approach to filtering junk e-mail. *AAI Workshop on Learning for Text Categorization*, July 1998, Madison, Wisconsin. AAAI Technical Report WS-98-05.

Sparck Jones, K. 1998. Automatic summarising: factors and directions. In Mani, I. and M. Maybury (eds.) *Advances in automatic text summarization*. Cambridge, MA: MIT Press.

Witbrock, Michael and Vibhu Mittal. 1999. Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries. *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR '99)*. 315-316.