Chapter 10.
Human Language Technologies

Figure 10.1: Illuminated manuscript (from Leige).

Here, we consider a variety of language processing applications. This includes text and speech processing and it looks at several specific applications such as text processing for search and question answering. Some ways of structuring multimedia may also be considered as language-like; these are examined in the next chapter.

10.1. Text

Text is an extremely flexible mechanism for transmitting knowledge. It is a fixed visual form that can appear in an almost infinite variety of locations, from books and magazines to road signs and T-shirts. Text, as a means of communication, is woven together much like a fabric is woven, with multiple individual threads joined tightly together to form a unified final whole; in fact, the word "text" is derived from the same root as the word "textile". Text is the visual form of a representational language. In addition to the shape of the text itself, the visual factors of how text is presented can indicate structure and convey meaning. Text is a common, fixed, relatively permanent and easily processed. While European-based languages use alphabets, other languages use syllable-based systems. Discrete versus statistical representations.

10.1.1. Alphabets and Character Sets

Historically, the letters in a alphabet represented the distinct sounds – the phonemes – of a language. Beyond the alphabet, other symbols associated with a language such as digits and punctuation are known as graphemes. Symbols.

The Latin Alphabet and the ACSII Code

There are many writing systems and many ways in which text symbols may represent concepts. English is written using the Latin alphabet, which is familiar to Westerners. Alphabets are sets of discrete symbols that are combined to form words, but individual letters (generally) have no meaning in themselves; it is only through their combination with other letters that meaning emerges. The 8-bit ASCII codes maps to the Latin alphabet allowing it to be recreated. While there are only 26 letters in the English alphabet, separate codes must be given to capital letters as well as the many punctuation marks and special characters used in writing.

Syllable Languages and Ideogram Languages

The written form of several East Asian languages, such as Chinese, Japanese Kanji, and Korean, uses pictorial ideograms to represent words (Fig. 10.2). Ideogram-based languages are not easily automated using current technology. Keyboard entry of ideograms is difficult because there is a such a large number of different characters. Compositionality of Chinese characters.

Figure 10.2: Part of the label on a package of Chinese herbal tea. Translated literally from top to bottom the Traditional characters say: (1) 10,000, (2) applications, (3) tasty, (4) harmony, (5) tea. (check permission)

Because of these difficulties and because of the need to assimilate foreign words, many Eastern countries have adopted phonetic alphabets. The Japanese use a mixture of traditional ideograms, a phonetic alphabet known as Katakana and some of the Latin alphabet.

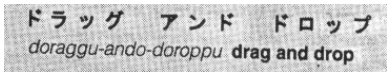

Figure 10.3: Example of a Katakana phrase [?]. (to be typeset)

There aren't enough bits in the eight-bit ASCII code to represent the symbols used in all languages. The limitations of ASCII are most serious for international character sets. Even some common symbols from European languages, such as German umlauts (e.g., ü) and French cedillas (e.g., ç), cannot be represented. Many proposals have been made for representing other non-Latin character sets. The Unicode standard was developed to extend the number of characters that can be represented and to facilitate interoperability across information systems. Unicode for sign language.

Ordering and Categorization of Characters

When we learn to recite the alphabet, in addition to learning the letters we also learn a standard order for them. This ordering is essential for looking up words in a structured list such as a dictionary, card catalog, or index (2.5.3). Traditional Chinese characters are categorized by "radicals". The characters for "flower" (Fig. 10.4) and "grass" share symbols to indicate that both are related to plants. This is the root form of the word and it is the basis of dictionary ordering. Once the radical has been identified, ordering is determined by the number of additional brush strokes required to draw the character.
indexradical, Chinese characters

草 花

Figure 10.4: The Chinese characters for "grass" (left) and "flower" (right) both share the "radical" at the top indicating that they are related to plants. Indeed, the radical resembles a plant growing out of the ground.

Orthography

The rules and customs for written text are known as "orthography". Orthography enhances readability but may also add aesthetics. Calligraphy.

Graphic design includes both text and image. Motion graphics animates graphic design. The style, or look, of a set of characters is its font. Most of this book is prepared with two main font styles The main text is in Roman font, in which the characters have "serifs" (short lines extending from the letters), while the captions and headings are in Helvetica, a sans-serif (without serifs) font. Fonts provide a regularity that makes the text pleasing to the eye. The spacing of the characters may also be manipulated to improve readability. Adjacent characters may be fused or split by a technique called "kerning" (Fig. 10.6). Font design and humanisitc fonts (Fig. 10.7).

Just as the inflection of spoken words can highlight the intended meaning written text contains clues about the meaning. The convention of capitalizing proper names allows the reader to identify them

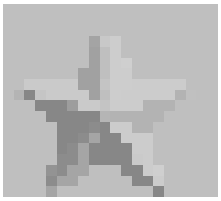

Figure 10.5: An example of Arabic calligraphy. Calligraphy was said to be one of the inspirations for Steve Jobs' emphasis on design.

RAVEN RAVEN

Figure 10.6: Spacing can greatly improve readability with typeset characters. Note the gap between the "A" and "V" on the left; no kerning has been applied. On the right, the letters "AV" have been kerned, in other words they have been moved closer together.

Figure 10.7: Fonts can improve readability. Clearfont font (dark "a") has recently been developed to improve the readability of highway signs compared to the older font (light "a").

more easily. Some other examples of orthography are the spacing between words, indentation between lines in an outline to show structure, and the ubiquitous smiley face :) emoticon.

10.1.2. Text Document Layout

From text to text documents. Equation layout (9.7.2).

10.1.3. Information Architecture

Web sites need to be managed. Consistency in design and navigation across the pages of a web site. Labels. Generalize principles of authority control across sites. Facilitate semantic publishing. Content management systems.

10.1.4. Digitization

Scanning. Digital photography. Color and image processing. Eliminating distortion from scanned image with post processing by de-warping. Large amounts of paper and text. Mass digitization (10.1.6).

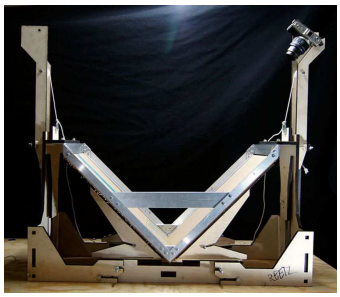

Figure 10.8: This book scanner is designed so the page images can be captured without breaking the spine of the book. (check permission)

10.1.5. Recognition of Printed Text and Handwriting

There are many advantages in converting paper text to electronic text. Because this is essentially visual processing, we need an image to work from. In some cases, only bitmap images of documents are available, rather than the full text. The electronic version of a document has been lost, or may never have existed. Those lost versions may be recovered by scanning (-A.18.2).

Optical Character Recognition (OCR)

OCR can be a step in citation extraction and indexing of digitized text. [26]. Optical Character Recognition (OCR) processes the visual properties of text in a scanned image of a page. It can also be used in detecting text in images or video which is a type of image-recognition problem. The quality of the original production affects the quality of the characters; in poor-quality printing the characters are often irregular (Fig. 10.9).

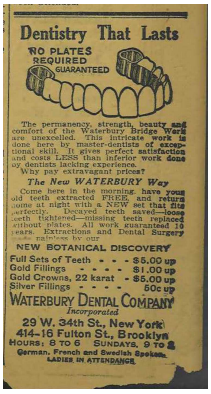

Figure 10.9: A scanned historical text. Note the uneven quality of some of the text. This adds to the difficulty of OCR.

The quality of the digitized image affects the OCR system's ability to recognize characters. As with other recognition processes, we may consider template-based and feature-based approaches. As a bottom-up model, OCR seems straightforward; individual printed characters are identified and combined into words. Individual letters may be identified by shape and attempt to determine words from them. Some of the errors that arise can be handled with spelling checkers (10.4.1). Unfortunately, this is far too simple a model. At the word level, some incorrectly identified characters may produce legal words while at the character level, the total number of characters may not even be correctly counted.

In complex recognition processes, some recognition occurs "bottom up". Bottom-up processing consists of gathering information from many sources and making an inference based on it (left side of Fig. 10.10). In speech recognition systems, an optical character-recognition (OCR) system might work bottom-up by identifying letters and constructing words and sentences from them. Other recognition systems work "top-down". Top-down processing starts with expectations about what is likely to be found; subsequent observations are used to confirm or reject those expectations (right side of Fig. 10.10). Characters might be more accurately identified by considering the possible words in which they may occur. The combination of bottom-up and top-down processing is known as "up-down" processing. In the previous example, the likely words would depend on the range of sentences appearing in the document in which they occur, effectively merging the two processing strategies. OCR performance can be improved by considering the lexicon and other linguistic constraints.

Processing OCR can be an example of parallel up-down processing (Fig. 10.11) (Fig. 10.12).

Much of the automated OCR for historical documents is poor because the print quality of the original documents is uneven. Crowdsourcing can be applied to OCR correction. Fig. 10.13 and Fig. 10.14. [6].

There are some additional difficulties in text recognition, such as segmentation of characters in Thai

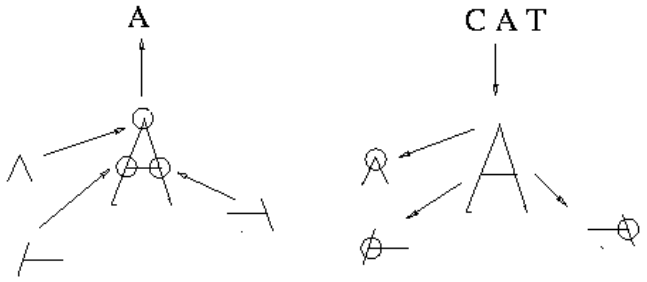

Figure 10.10: Two approaches for visual recognizing of the letter “A”. In the “bottom-up” approach (left), the model is confirmed if a sufficient number of features are matched. In a “top-down” approach (right) the context provides expectations that can be confirmed by looking for specific features. (redraw)

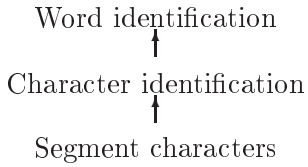

Figure 10.11: A bottom-up approach for OCR.

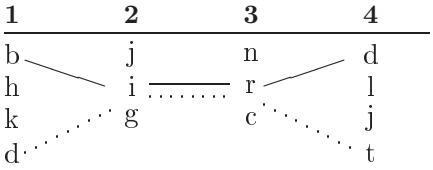

Figure 10.12: Characters may not be identified exactly. Thus, they may be ranked by probability and then the word from a word list with the highest net probability could be found. The word “bird” is preferred over “dirt”. (redraw) (under construction)

Figure 10.13: Text-correctors Hall of Fame from the Australian National Library[9]. (check permission)

(Fig. 10.15). Optical processing can go beyond recognition of text to the identification of other symbols, including non-traditional characters, musical notation, and mathematical formulas.

Visual Document Segmentation and Recognition

Uses of document the document are reflected in its structure (2.3.3).

A complex document, such as a magazine, has several different page styles; for example, there are title pages, tables-of-contents pages, and text pages (Fig. 10.16). Moreover, the text pages may have additional structure such as section headings and figures. A first step could be the identification of the document. This is a type of visual language. Determining visual structure can be useful for information

Figure 10.14: reCAPTCHA. (check permission)

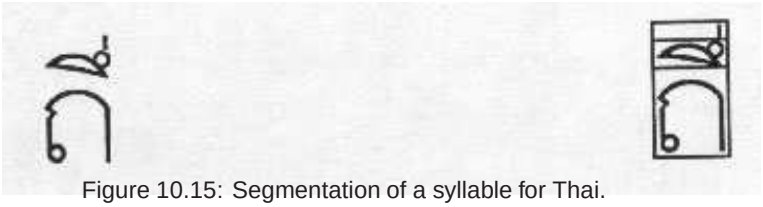

Figure 10.15: Segmentation of a syllable for Thai.

extraction.

Figure 10.16: Despite the apparent visual similarity; a document detection system should be able to determine that the first two of these pages are title pages, while the third is not. (check permission)

The recognition of the content from scanned documents is a special case of image segmentation and image processing, which we will examine in more detail later (-A.2.3). Visual document processing might begin by determining the layout, such as whether the text was printed in one column or two and the positions of the figures and their captions. Fig. 10.17 shows a newspaper which has been processed to identify text zones. Visual prototypes and deformable models can be used for page layout. Just as the shape of individual objects can be recognized, the layout of a whole page might be parsed as a type of visual language (11.2.4).

Handwriting Recognition

Handwriting has much more variability than does printing. To begin with, handwritten letters are sometimes block and sometimes connected, or cursive. Static Handwriting. Recognize content in manuscripts. Fig. shows a sample of handwriting. Some features such as the loops (l), ascenders (t), and descenders (y) stand out clearly. Thus, the first step would often be the segmentation of individual letters.

Dynamic Handwriting. The sequence of pen strokes in producing printing and handwriting can be used to improve recognition. These techniques could also be useful for signature verification. Fig. 10.18 shows an example of how this could help to distinguish between a "U" and a "V". For ideograms character recognition, stroke order is particularly helpful. The recognition of the pattern of strokes is like other sequence-recognition problems such as the recognition of 2-D gestures (11.4.1) and HMMs. Pen-based interface.

Symbol recognition. Many more symbols than but more constrained patterns. This can employ dynamic

Figure 10.17: Zone segmentation for an OCR program for a historical newspaper image. This segmentation is based on the spaces around the blocks of text[32]

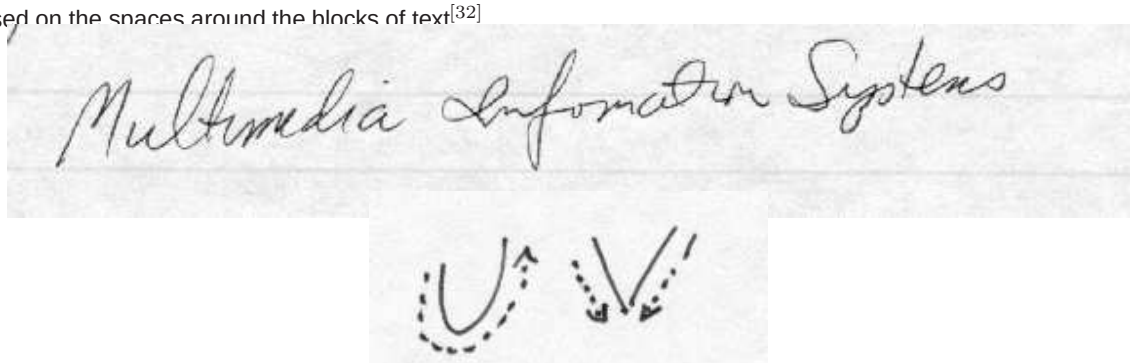

Figure 10.18: Although the printed letter U looks similar to the printed letter V, they can be easily distinguished by the strokes used to write them, as shown by the dotted lines. (to be rendered)

OCR (10.1.5). Operations on math (Fig. 10.19).

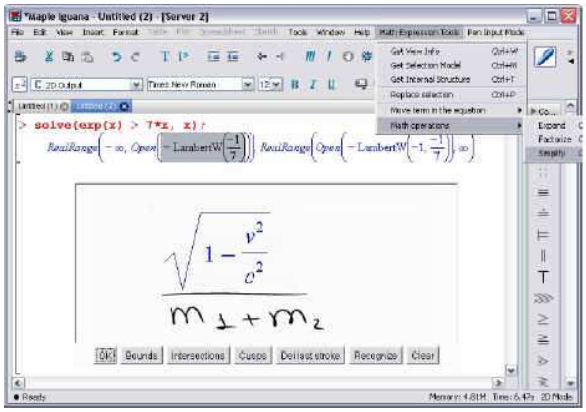

Figure 10.19: Pen-based input is effective for mathematics [?]. (generate own version) (check permission)

10.1.6. Digitized Collections of Books and Documents

Until recently information resources were committed to traditional media. Supporting collections this traditional material. This content can be digitized and can be organized into collections coordinating paper libraries. Digitization can facilitate preservation, distribution, and access of fragile material.

Because there is so much paper. Building the collection. NDNP. Millions of digitized pages of newspaper text will be available online soon. Million Book Project[54]. Multimedia. Google book scanning project. Digitization for access. Digitization workflow processing.

Consistent with the theme of interactivity Digitization for access rather than for preservation. Spatial relationships can contribute to meaningfulness. Visual languages (11.2.4) and image segmentation.

Much historically important material is found in manuscripts[17] and other personal papers — these writings often reveal the innermost thoughts of their authors. They are often kept in archives (7.5.1). Manuscripts and other original documents are "primary sources" and are often used teaching history. However, primary sources have limitations, In personal letters and memories people may puff up or simply lie about their contributions. Issue of authenticity of primary sources. Their interpretation can be difficult without the context.

In some cases, texts are digitized because they are deteriorating. Moreover, the resulting image can be processed. That is, electronic restored.

Printing has been used to produce books for hundreds of years, but historically, it was employed primarily to prepare only publications that were meant for a wide circulation. This means that over the centuries a vast archive of handwritten documents has accumulated. Until about 1900, typewriters were not often used for business correspondence and until about 1950, almost all personal correspondence was handwritten.

Collections of manuscripts are thus rich sources of historical material. Many of the original manuscripts of Mark Twain's books are kept in the Library of Congress, as are a number of original music scores. These early versions may include author or composer notes, which provide a fascinating insight into the work and/or creator, but which do not appear in the final publication.

Collections of "e-text" often combine traditional and digital approaches to collection management. The scanned images may be OCR'd and metadata may be extracted via the document processing and recognition techniques described above (10.1.6).

Semi-structured text. For instance, email has structured fields and free text.

The levels of abstraction are applicable beyond defining metadata. Instances of documents may appear in several different media and in different formats, such as the derivative works mentioned above, or even audio recordings of printed works such as books-on-tape[68]. A document viewer that allows several versions to be presented simultaneously (Fig. 10.20). An original text might be presented along with the output of optical character recognition (10.1.5). Document browsing. Multi-layer documents. Image based-electronic editions.

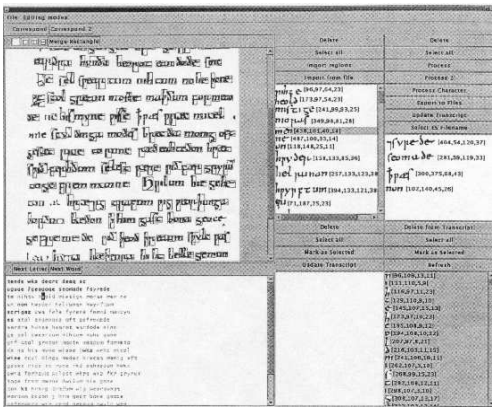

Figure 10.20: Multiple versions of documents can be compared through a single interface[33]. (check permission)

Increasingly, versions of a work proliferate. Levels of context to include. Metadata.

Superworks and families of works. Equivalent, derivative, and descriptive.

Ecolosystems of related documents. Annotations, Reviews.

Extracting Metadata from Text Documents

The next step would be to extract the content of the document. This can be done by processing the structure of documents. Fig. 10.21 shows an example of image extraction. The middle image shows parts of the page which have been confidently identified as not being text regions. Because so many of those regions are clustered together, it seems likely that the whole region is a picture. There is a similar challenge in extraction of values in a table can help one understand the structure and semantics of the table. Styles of tables.

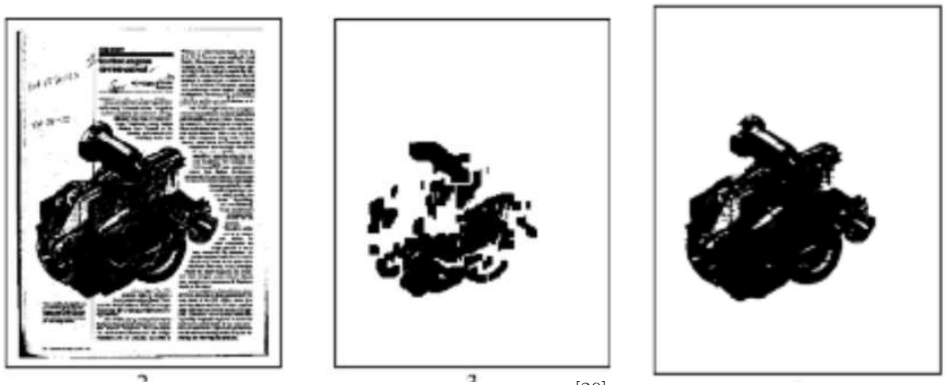

Figure 10.21: Stages in image extraction from a magazine page[20]. (new version) (check permission)

Once the title page has been identified, metadata from that page can be extracted to populate a catalog for an e-text database. That metadata can facilitate document classification, which could be useful for automatically creating a style-sheet for the document or metadata for a collection of documents.

Massive Digital Collections

Mass digitization.

Digital humanities (9.1.1). Libraries and archives (7.5.1). In the past, information systems have focused on business and natural science. Increasingly, information systems are begin used in the humanities. Mining digital humanities materials.

Literary criticism. Background of the author as reflected in the work.

History, literature, art, archaeology.

Google Book Project. Copyright issues (8.2.2).

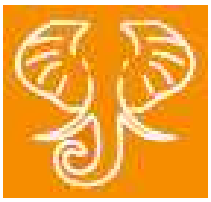

Figure 10.22: The Hathi Trust (logo shown here) manages a very large collection of digitized books. (check permission)

Digitization of non-text. 3-D digitization ((sec:3Ddigitization)). Cultural informatics.

10.2. Reading

Ultimately, the purpose of a text is to convey information to the people who read it. At the simplest level reading is just the process of extracting meaning from writing. The technological developments of recent decades have had a significant impact on the acts of reading and writing. For the user, writing with a word processor is different from writing with paper and pencil. Writing and reading are complex human activities which have both cognitive (4.3.0) and social foundations. Education (5.11.0), and libraries (7.2.1). Written language is often more formal than spoken language (11.3.3). Discourse (6.3.2). Reading as constructing meaning. This is related to playing games (11.7.0).

The experience of reading produces a number of cognitive psychological effects on the reader [?, ?]. While the subtlety of an argument cannot be evaluated automatically, readability is affected by the complexity of the words and syntax in a text. Reading requires human language understanding. Because language is redundant, it is possible for somebody with a weak vocabulary to augment their understanding by deeper analysis of the structural relationships among the terms.

From reading to storytelling (6.3.6) and orality (11.3.3). Reading technologies (5.11.5).

10.2.1. Reading Documents, Books, and Hypertexts

While we think of reading a sitting quietly with a book, in fact reading is a highly varied activity. Books (8.13.6).

Figure 10.23: Types of reading: A) boy reading, B) nutrition label, C) poetry reading, D) Social reading. (check permissions)

Reading is a complex cognitive activity. Scholarly reading environments. Reading aids. Long form presentation of arguments in books.

Hypertext increases usability of online material and allows the user many choices, but this can potentially lead to discontinuity as the reader moves from one topic to another. Often media are only loosely coupled. and there is a lack of continuity – some hypertext contents are fragmented, others are structured. It is helpful if hints are included about where links will take you. On one hand, there is a distraction from clicking and jumping to different topics but there is also an advantage of being able to move quickly to topic of primary interest.

Jumping into the middle of a document without a coherent introduction can be disorienting. Part of reading is understanding the structure of books. Readability and document layout structure such as page numbers. News headlines, lead with the main news.

10.2.2. (Reading) Literacy

Reading has long been regarded as important for active participation in society — so reading skills impact not only the individual who acquires them, but also society itself. Reading allows individuals to learn about their society's conventions, laws, culture, etc. Reading as integral to scholarship and critical thinking.

Assessing literacy in children. Technology and learning to read (5.11.5). Increasingly, the various literacies are converging. Practice in reading directly affects literacy.

Social factors affect the ability to access written information. Reading is often associated with critical thinking. Because the new information technology has vastly increased the ease with which images and

audio can be disseminated, literacy may become less common, with speech communication becoming as general as written communication. We may move from the symbolic reasoning that results from reading text to thinking more pictorially and orally.

10.2.3. Reading Comprehension

Reading is more than processing individual words and sentences; it involves making inferences about the material being read. Fig. 10.24 shows a sample reading comprehension test. This task can also be used to evaluate the performance of automatic question-answering systems (10.12.0).

Discourse comprehension. Discourse elements and text cohesion (6.3.2). In many cases, reading comprehension is related to sensemaking.

> Up to 40% of heating and cooling needs in the home are due to air exchange. In the winter, heat losses occur when heated air leaks out of the house. Energy must then be spent warming the cool air that replaces it. In the summer, heat gains occur when warm air leaks into the house; thus, more money is then spent on cooling costs.
>
> One area of the house that may allow up to 17% of this unwanted air infiltration is around windows and doors. Energy waste can be cut by the proper installation of windows and doors.
>
> Which of the following can be inferred from the above description?
> 1. Air infiltration around windows and doors adds about $40 to each month's utility bill.
> 2. Homeowners can save on their utility bills by reducing infiltration around windows and doors.
> 3. Additional air infiltration occurs around chimneys, electrical outlets, and light switches.
> 4. Infiltration is a problem in both the summer and the winter.

Figure 10.24: Sample reading comprehension task (adapted from[3]).

As a person reads, there is an interplay between low-level perceptual processes and higher-level cognition. For instance, excessive eye fixation on a particular passage suggests that it is difficult to understand.

Reading comprehension and expectations based on semantics, syntax and genre, and norms. Bartlett (4.4.3). Thematic organization. Exposition (6.3.3).

Importance of reader's background on understanding narrative (6.3.6).

Certain discourse structures can ease comprehension[16].

Macro-rules for reading and summarization (2.5.5, 10.6.2)[34]. Conceptual synthesis and inferences made while reading.

10.2.4. Skills Needed for Reading

Reading is not a unified, simple activity. Rather, people interact with documents in many ways. However, reading is not a simple or easily defined behavior. Let us consider how people interact with text documents. Reading for participation in a text-oriented society. We regularly use reading to acquire information[10]; it is also a source of great pleasure. Reading is a foundation for accessing information. Understanding conventions such as chapters and page numbers.

Information technology is changing the use of text. We have already discussed hypertexts and annotations. It is even possible to imagine the decline of reading and the increased use of verbal language because of multimedia systems and speech-recognition technology. Reading and linking. Can we improve on paper books.

Reading is probably best thought of as a set of skills which must be mastered[49]. We learn, for example, the direction of the text. We also become able to associate printed words with their corresponding

sounds; phonemes and phonics are two tools used to acquire this skill. Phonics. Difficulty of the English alphabet code[44].

Eye fixations are shown for a simple text passage in Fig. 10.25. In addition to the words themselves, many factors such as orthography (10.1.1) and layout structure affect reading.

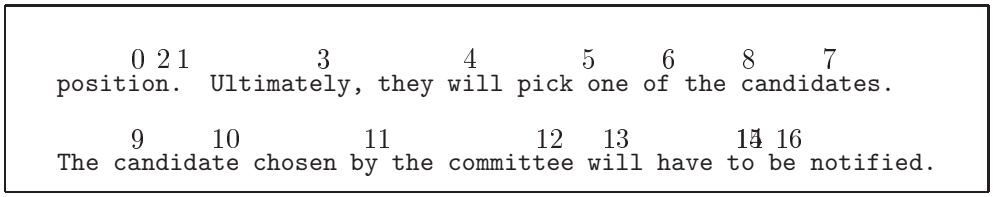

Figure 10.25: Sequence of eye fixations while reading. Note the fixations on the most important words in the text. Also note that the order is not perfectly sequential[1]. (check permission)

Cognitive Availability of Text

Orthography (10.1.1) can improve the readability on a display. Fonts. Layout to emphasize meaning. Pull-outs. Fonts. Computer displays are often poor devices for reading. The fact that they are back-lit is one problem; another is the difficulty of drawing truly straight lines on a typical computer display. Online reading also involves scrolling, which some may find tedious.

Electronic presentations may have fewer external clues to provide reference points as memory aids compared to paper.

Well-written text should match the reading ability of the students for whom it is meant, for example, in the complexity of its vocabulary and structure.

10.2.5. Close Reading and Active Reading

There are many forms of reading ranging from quick scanning to "close reading". Pre-reading. Close reading is trying to understand in detail what the author intended and how the author accomplished those goals. Transliteracy. Disputed texts.

Textuality versus reader-response theory and conversational analysis. A poem is "what the reader lives through under the guidance of the text" [57]. Intertextuality compares texts and considers how they relate to each other.

Reader Response Theory. Literary reception theory. How much of the meaning of a work is in the work itself.

Active reading is a way to interact with documents. Reading hypertext and "meta-reading" [50][??] Potentially, hypertext allows people to navigate to the parts of the text that t meet their information needs.

People often like to interact with the material they are reading[10], [51]. For instance, students may underline or highlight sentences and make notes in their texts. They are "active readers". Fig. 10.26 shows highlighting added to an e-text (10.1.6).

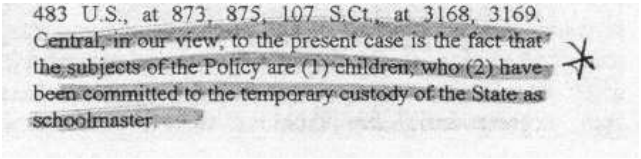
483 U.S., at 873, 875, 107 S.Ct., at 3168, 3169.
Central, in our view, to the present case is the fact that the subjects of the Policy are (1) children, who (2) have been committed to the temporary custody of the State as schoolmaster

Figure 10.26: Highlighting is one way to personalize an information resource[27]. (re-create) (check permission)

Reading requires continual integration of new information with a user's existing knowledge; this in-

tegration can be facilitated by reading aids that encourage assimilation of and reflection about the material. Titles, marginal notes, and structure markers may serve as "cognitive organizers" for readers [?]. Ultimately, this is a type of sensemaking. Cognitive organizers can also be graphical.

10.3. Writing and Text Communication

10.3.1. Writing and Authoring Text

Writing involves many types of cognitive activities, from motor responses to language generation. Indeed, writing may facilitate the assimilation of knowledge, as note-taking forces deeper processing of the material (4.3.3). The full process of composition involves language generation such as planning and revision. Writing is often done without fully knowing the audience. Writing as an ongoing content development process. Language generation (10.4.3). Information extraction (10.5.0).

Rhetoric in persuasive writing. Authoring multimedia (e.g., film). Argumentation (6.3.5). Constructing texts with complex argumentation[63] Argumentation zoning.

Mechanics of Writing and Editing

Writing combines complex motor responses with complex cognitive processes in composition. We discussed handwriting recognition (10.1.5). Handwriting and keystroke production are complex examples of motor behavior (4.2.4). Errors in typing may be due to motor or cognitive difficulties. Many of the errors are cross-hand transpositions such as typing "whihc" instead of "which". Word processing has changed the way writing is done.

Text entry on mobile phones.

Many alternative keyboards have been proposed as well as other text input modes such as pen-based entry and editing.

Authoring Text

Composing text is one type of language generation. Just as graphic design emphasizes using visual media to convey a message effectively, design can be applied to natural language to convey meaning and emotions, and to enhance clarity. This design allows for statements to be structured effectively, and for the construction of topics sentences. Language generation (10.4.3). Discourse types (6.3.2).

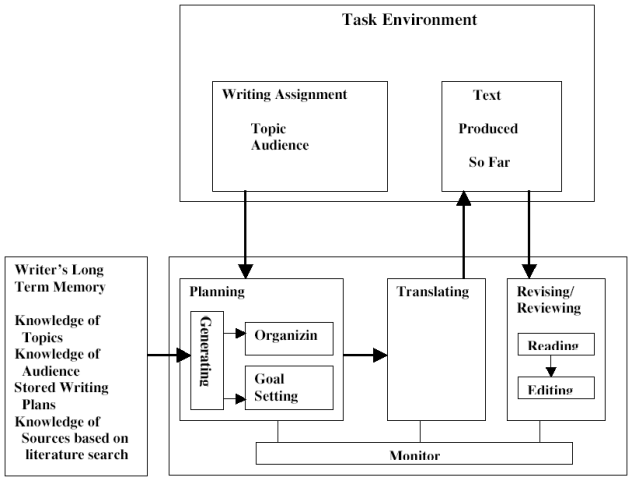

Figure 10.27: Schematic of writing and editing process (from[28]). The process shown here may be repeated through many iterations. (redraw)(check permission)

Writing and new media. More versions and the management and collection of those versions. Digital lives of authors. Cultural preservation.

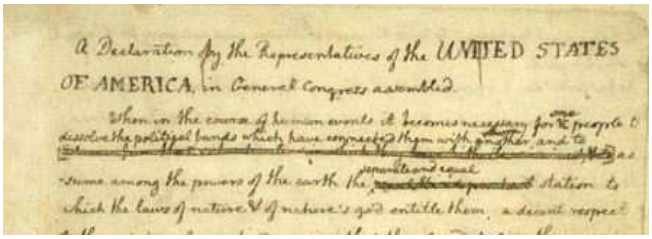

Figure 10.28: Here are edits to a draft of the Declaration of Independence. Saving the digital versions of an author or artist. (different example) (check permission)

Assessing Readability and the Quality of Student Essays

Student essays teaching reading. Perhaps surprisingly, statistics can help in assessing good writing. Good writing should be easily readable, and there are two easily measured factors that contribute to the readability of a text: the number of words/sentence and the number of syllables/word[61]. These are combined as shown in Eq. 10.1 to determine a Reading Ease score. Text coherence.

$$Reading\ Ease = 1000 - (5 \times \frac{words}{sentence}) - (400 \times \frac{syllables}{word}) \quad (10.1)$$

Reading Ease is one component of essay-grading scores (-A.6.4). The reduction of readability to such a simple formula as this is frequently criticized. Style, layout, and font also have an effect on readability. Student writing is highly variable, and it is often difficult for teachers to assess it fairly. In some cases such as writing samples on a large standardized test, tens of thousands of essays may need to be scored with as high a degree of reliability as possible. Some standardized tests are graded by one human reader and one automated reader. One strategy for automatic grading employs information extraction would look for high counts of words such as "because" that suggest the author is giving evidence for his or her arguments; this is seen as a sign of good writing. Fig. 10.29 shows two paragraphs rated by a technique discussed in (10.3.1). (-A.6.4)[22].

> The human heart is divided into two parts: left and right. In infants this division of the heart is not so clear as it is in adults; although by the age of one the division becomes clearer. The heart receives blood, and pumps it to the other organs in the body, so they can perform their functions. Most important the heart pumps blood to the brain so it can dictate to the other parts their daily functions...

> The heart is the main pump in the body that supplies the rest of the body with oxygenated blood by way of the arteries. There is a reaction that takes place at the ends of the arteries where the oxygen is taken out of the blood and replaced with CO_2. This reaction takes place in the capillaries. The arteries branch to become smaller and more numerous, these are then called arterioles...

Figure 10.29: Excerpts from two essays about the circulatory system which were graded by Latent Semantic Analysis. The first excerpt was given a low score (1 out of 5) while the second excerpt was given a high score (4 out of 5)[37]. (check permission)

10.3.2. Text Communication and Collaboration by Text

Text messages are widely used for communication between people. Because of bandwidth limitations and computer capabilities, interaction via text is still very common and effective. Because transmission of text is cheap, it is widely used, and supports a wide variety of collaborative services such as online chat, emails, newsgroups, online chat is an interactive, live, text discussion.

Communication modalities (5.6.5).

Social interaction mediated via text. Creating online communities (5.8.2). Collaborative tasks and facilitating coordination. Conversation (6.4.0).

Telephony and connectivity (5.1.5).

Email

y Unlike chat or live telephone calls, email is asynchronous. That is, once an email message is sent, it can be stored until the recipient is ready to access it. Email readers also receive metadata such as the email address of the sender and the date it was sent. Email introduces new features such as replying to messages while appending previous messages; thus, a thread is created that provides a context for the message. Because of its flexibility, email is more effective at facilitating social interaction than many tools which have been developed specifically for that purpose.

Email is an asynchronous medium. Unlike synchronous interaction, the synchronous interaction in email includes reflection about the dynamics of the interaction.

Texting. Mobile service.

IR for non-standard content (10.11.1). For instance, email files can become archives and can be searched. Fig. 10.31 shows a fragment of the email protocol. This involves aspects such as linking of email files, finding conversation threads, and determining social networks (5.1.0). Using roles and honorifics in determining group structure from email archives.

> Correcting HTML seems a pretty hard problem. Tidy makes a
> valiant effort and gives us a good start on the job in many
> instances. Doing everything might be more than can be hoped for.
> Try making a long outline in MSWord, save it as HTML, and then
> try to clean that up.
I succeeded in being able to infer a noframes element and in
moving the trailing body inside the noframes, and in placing
the new noframes within the outer frameset. The fix is part of
the April 15th release.

Figure 10.30: Sample email exchange (from TREC). (check permission)

Emails and social norms.

Email spoofing. Email mining. Privacy records.

Gmail and advertising. Increasingly, spam leads to security attacks.

Filtering Out Spam Filtering can also be used to block the presentation of unwanted material such as spam. It is difficult because the topic of the spam is not known. However, the accuracy of these filters is often low; for instance, filters may be tuned only for English. One of the most successful techniques is filtering based on linking (10.10.2); for instance, sites containing pornographic content are often linked together. This may be used to filter spam. Among the specific methods which have been proposed are communications procedures, Bayesian filtering (-A.8.2), and collaborative filtering. It's harder to determine what the user will consider junk.

Two methods. User authentication. Sender bond which is something like a stamp. Essentially, this is the reputation of the sender (5.2.2).

1ow cost 1nsuranCe
personal Medic@ti0n

Figure 10.31: Examples of mis-spellings which are indications of spam.

About 90% of email traffic is spam[7]. Because it is a free good, it is abused. A spam filter could be based on text processing principles. Filtering out (10.3.2). Another alternative would be to change the nature of the email service. A service could require a "stamp". Another alternative would require a valid return address for a message to be valid. Blacklist.

Newsgroups and Threaded Discussions

Extended and ongoing discussions may help to form a community (5.8.2), as a discussion might evolve from threads to concepts to topic-oriented threads. Discussion groups often branch off into very different topics and it would be helpful to have tools for keeping the "threads" of these discussions straight. This may be seen as an evolving hypertext, although threading gives it topical continuity. Structured conversation.

Roles for a group discussion; a moderator may manage, focus, and re-conceptualize topics to facilitate discussion. A "lurker" will observe a discussion without contributing to it. Newsgroups are a favorite haunt of lurkers. A "newbie" is somebody who is new to a group. Social networks (5.1.0).

Netnews as a content archive, DejaNews.

Indexing threaded discussions.

Supporting Collaborative Interaction with Information

We have seen many forms of collaboration, Collaborative organization (5.7.0) and communities (5.8.2). This is one type of collaborative task and groupware environment (5.6.2). An activity associated with organizational sense making (5.7.3).

Tools to support collaborative search. Shared retrieval versus collaborative understanding. Information commons.

Collaborative creation of knowledge. Collect information from a range of people[66]. This is an aspect of Web 2.0.

We have considered question answering in many places in this text. Get people to answer questions like reference service (3.3.2). Yahoo answers. Difficult of spam in responses and picking best one. Characteristics of users.

Encyclopedia. Wikis have proven effective as a platform for communities to organize and discuss issues. Wikipedia as a organizing system. Very good for some topics. Probably less accurate where somebody has a vested interest in certain perceptions. Possibility of a concerted attack on Wikipedia. Contrast this to traditional scholarly communication (9.1.1). Particular emphasis on popular culture.

Wikipedia is the community-edited encyclopedia. Wikipedia facilitates a neutral viewpoint by allowing open commentary. It is disputed how well that works in practice. It has evolved a number of governance mechanisms for ensuring quality[23]. Among the principles are that contributions are verifiable and that no original research is published. On the other hand, issues of the reliability of information ((sec:informationqulaity)). Criteria for articles: notability. Verifiability. Responsibility of a individual for the credibility of the information. Social transparency.

Wikipedia and as an online volunteer community (5.8.2).

10.4. Natural Language Processing

Language processing has many applications. Given the ubiquity of natural language we may consider how to process it automatically. Text retrieval and many other tasks may be considered as natural language processing. In text and natural language processing, shallow methods may be distinguished from deep methods. Shallow methods may include rules of thumb; deeper methods include full parsing. Speech processing and linguistics (11.3.3).

To what extent can the text be processed in a way that preserves meaning. The traditional model of language suggests it can be decomposed into several levels: words, phrases, sentences. There are many levels of language processing. Here, we consider processing for basic components such as words and then consider tasks that involve processing entire documents. We focus on methods appropriate to this breakdown; later, we consider statistical approaches to text-processing.

Figure 10.32: A portion of the log of editorial discussion on the Wikipedia article on the Boxer Rebellion (accessed on Sept 1, 2007). Note the editorial comments inserted into the head of the article. (check permission)

Lexical resource. Corpus linguistics.

10.4.1. Word-Level Processing

Spell Checking and Correction

A simple spell checker looks through all the words in a document and matches them against words in a dictionary. If a mismatch is found, the program may generate a word that may be what was intended. Several algorithms have been developed for determining the similarity of two words. Suppose a misspelled word were found in a document; candidates in the dictionary for the correct word are proposed to replace the misspelled one. The misspelling might come from a missing letter, an extra letter, or transposed letters. Suggestions for the correct word could be obtained from a calculation similar to the edit-distance calculation below. Suppose a person had typed "INFORVATION" into a word processor. The spell check program might compare that word to its dictionary. The easiest way to make this comparison is with "approximate string matching". If the dictionary contained the terms "INFORMATION" and "INNOVATION," the steps for finding edit distance are shown in Fig. 10.33. Several constraints can be added to these simple word-distance measures. For instance, we might weight possible spelling errors by the likelihood of their occurrence.

Steps	Change	Cost	Current Guess
0	original		INFORVATION
1	V->M	2	INFORMATION

0	original		INFORVATION
1	F->N	2	INNORVATION
2	drop R	1	INNOVATION

Figure 10.33: Finding edit distances between an incorrectly spelled word and two candidates from the dictionary. The term "information" would be selected as the match because fewer steps are required to match it with the incorrectly spelled word.

The second phase is spelling correction such as for queries submitted to a search engine. Specialized situations. Learning spelling from previous example. Spell correction by learning from web queries. Examples of the mistakes some people make and how they are corrected. Based on expectations. Spell correction based on examples.

Word-Sense Disambiguation

Word-sense disambiguation determines which sense of a word is meant in a given context. People can

often distinguish between word senses with a very brief context of surrounding words. One word sense in any given discourse. Many attempts have been made to automate sense disambiguation.

Word Segmentation

Earlier we discussed segmentation in OCR (10.1.5). For some languages, words run together when they are printed and this requires segmentation. An example from German would be *weltanschaung* or world-view — this is the perspective a person brings to their interpretation of events — where "welt" means "world" and "anschang" means "view" Dynamic programming can also be used for effective segmentation.

Part-of-Speech (POS) Taggers

As we noted earlier, parts of speech define highly structured language. Part-of-speech tagging is often used for language checkers such as those found in word processors; they are also useful for determining phrases (6.2.2). A simple, but inefficient, approach checks all words in a dictionary for all possible parts of speech that they can assume. Ad hoc rules could then be applied to improve on the original guesses [15].

10.4.2. Syntax and Parsing Natural Language

We have described the importance of syntax in (6.2.2) and we have already discussed parsing of formal grammars (6.5.1). If we focus on the syntax of a statement to understand it. In the sentence: "The cotton sails on the wind". it is essential to know that "sails" is a verb rather than a noun. This can be determined because there is no other word in the sentence that is likely to be a verb. Generative models.

Grammars can be also implemented with augmented transition networks (ATNs) (6.5.1) though the notation is not as compact.

Rule-Based Models

Earlier, we examined parsing formal grammars (6.5.2). Determining grammatical structure is one of the most obvious strategies for language. Grammars are often assumed to be the representation for natural language. Parsing attempts to identify the components that underlay a sentence. Fig. -A.40 shows an example of re-write rules and a lexicon for a highly simplified sub-sample of English. A number of parsing algorithms have been developed some of them are summarized in (-A.5.4). There is a particular problem of parsing ambiguous example sentence the "man on the hill".

State machines (3.10.1). Garden path sentences. ?The man who hunts ducks out on weekends?. Uncertainty and NLP state models. Toward HMMs.

Statistical Models

State machines make transitions from one state to another. by the fulfillment of certain conditions. These machines can also be probabilistic; "weighted automata" have probabilistic transitions between states. Similarity rather than categorization.

10.4.3. Text Generation

How can a computer make appropriate statements about the world? Many tasks are closely related to language generation such as the generation of explanations, translations and summaries. Human language generation may be modeled on several of the same strategies used for planning (3.7.2) such as generating and testing means-ends analysis.

The problem of language generation may be broken into two parts. The first is generating a natural language description from a symbolic representation. For instance, we might generate the description of a computer program. Fig. 10.34 shows several stages for such a description[46]. The basic concepts with an ontology (2.2.2). the description is organized (discourse planner, Section 6.3.2), words are chosen (lexicalizer), and finally a surface generation is produced. Text generation to accomplish pragmatics, not just semantics.

Stage	Description
Plan Generator	Determine goal
Ontologizer	Decide what concepts need to be included.
Discourse Planner	Decide how the concepts should be assembled.
Lexicalizer	Select words.
Surface Generator	Develop coherent surface text from the selected words.

Figure 10.34: Stages in language generation when starting with a formal specification (adapted from[46]).

Often, relatively simple rules can support generation. One such rule would be to give a full description of a named-entity the first time it is introduced. In other cases, sentences can be combined; for instance, in Fig. 10.35, the two sentences in the first line can be combined as shown in the second line to eliminate redundancy. This is an example of the rule that creates a modifier at the beginning of a noun phrase [46]; in other words it creates "ellipsis".

Camilla Jones visited 31 patients. She is a doctor.
Doctor Camilla Jones visited 31 patients.

Figure 10.35: The surface generator might combine two sentences (top) into one (bottom).

One of the maxims of communication is that material be adapted to a situation and recipient or user. User models (4.10.2) can be applied; from some of these, base-rate expectations are established about how to respond to a user (e.g., an automatic system for selling train tickets) with appropriate level, word choice, and sentence complexity. Sometimes high-level organization is required. One needs to know not only about the recipient's knowledge but also about the environment in which that person is working; for example, whether tools are available and what they are.

Complex text generation shares some similarities to conversation because it takes into account the goals of the reader. Thus, it needs to consider both user models and task models. However, the audience is often less well-defined; traditionally text is more formal.

Unrestricted conversation development is still difficult. Successful language generation applications often build from well defined specifications. Ultimately, language generation should include speech generation (11.3.3) and even extend to the coordination of expressive modalities such as gestures, glances, or hesitations in speech.

10.5. Information Extraction from Text

A lot of information is stored in text documents; these can be documents that were originally created as text documents, or scanned images of older hard-copy files. While this information may technically be available for the gathering, it can be difficult to extract it in a systematic way. Information extraction methods attempt to find from within a complex document its most essential information, such as dates, numbers, names, etc. These few facts could be sufficient for simple question answering (10.12.0).

Extraction from semi-structured data sets. For instance, find a name or a date on a web page.

Simple text-processing methods[29]. This is useful for question answering (10.12.0), automatic metadata and semantic annotation assignment, summarization (10.6.2), and mining blogs.

Because many concepts are complex, with multiple elements contributing to an overall idea, it is difficult (currently impossible) to create a system that understands complete concepts. It is easier, however, to find concepts associated with specific terms. Knowledge-based extraction, though an improvement on techniques based on key words or numbers, is a still shallow method for extracting some information from text. It looks for specific patterns, either of individual words, or more complex, but regular, arrangements. These regular arrangements could be numbered instruction lists, for example, or directions — both of these things tend to take the same format, and could be searched for using knowledge extraction techniques.

Language has a lot of structure beyond syntax. For one thing, structured information allows a program (and those creating it) to form fixed expectations about what information and attributes a document will contain, where it will be contained, and how it will be presented. Many of these methods are "lightweight"; that is they do not attempt to analyze the meaning but use only surface clues (10.4.0).

Using templates.

Discovering new facts from a collection.

10.5.1. Named-Entity Extraction and Entity Resolution

Names provide generally distinct identifiers (2.2.1). They are helpful in understanding text such as a news story. Names in text have a variety of distinctive clues [47] that identify them. They are usually capitalized, and, grammatically speaking, are treated as nouns. A variety of techniques can be employed by a system when attempting to distinguish people's names from place names or other possible confusions, such as the use of "Mr." or "Mrs." or the proximity and semantics of certain word combinations. Multi-tiered and cross-referenced screening processes can provide strong evidence of the use and type of proper name (10.4.0). There are, however, difficulties in named-entity extraction techniques; Fig. 10.36 illustrates the fact that even names can be ambiguous.

1. Washington gave his Farewell Address.
2. I drove to Washington.
3. The boat was near Washington.

Figure 10.36: Is Washington a place or a person in these examples?

Finding all references to a given individual can be very difficult. Washington might also be referred to as: The "father of his country", the first President of the United States. Many names are in common: "James" and "Jane". Same-person with different names, different people with the same name. Named-entities and name authority (2.2.1). Social networks. Named-entity disambiguation.

Further, the specific type of named-entity can sometimes be identified from the surrounding context.

Semantic annotation is adding semantic descriptions to objects. These descriptions are often drawn from ontologies.

Because it is relatively robust, POS tagging (10.4.1), is a good place to start for named-entity extraction.

Named entity extraction as feature extraction.

Using a name authority for named entity extraction.

Lists of known names are useful for disambiguation.

Entity co-reference problem. Co-reference resolution.

Types of named entities. This is useful in question answering.

Creating an index of named entities. Using the Web to extract named entities because the Web is so large that there is a lot of redundancy.

10.5.2. Extracting Conceptual Relationships using Verbal Templates

Frames are structures for information. Some information can be extracted into a frame [8]. Frames and templates lead to document segmentation (10.1.5). This segmentation can be helpful for metadata extraction, but it can also cause concepts and information to be presented too simplistically, or even incorrectly Fig. 10.37 shows some examples of templates.

Structure may provide useful clues. A classified advertisement often has a predictable structure. Similarly, modeling the structure of a table is essential for extraction of content from those tables.

IF you go to the store **THEN** please buy some bread.
WWW **STANDS FOR** World Wide Web.
a car **IS A KIND OF** vehicle.
ON ONE HAND Jane was happy **ON THE OTHER HAND** Jill was sad

Figure 10.37: Linguistic templates can be mined to extract useful information. Here are some template markers highlighted with bold font.

For complex information, establishing what category is associated with each word and dealing with co-references between information categories is hard to do with complete accuracy. Much of a template's functionality is determined by its construction; there must be both flexibility and specificity to make information extraction possible or effective.

One way of accomplishing this information categorization is to apply statistical analysis to information to establish the parameters, or rules, by which the observable content is defined. While there are different methods, Hidden Markov Models (HMM) are particularly apt in that they work backward, deriving the underlying parameters from an aggregate of data that display visible patterns. These findings can then be fine-tuned for further accuracy.

10.5.3. Applications of Information Extraction from Text

Factoid extraction for question answering (10.12.0). Consider the statement "Marie Antoinette was the last Queen of France". This may seem obvious enough but consider the wide range of knowledge it implies and the additional facts that are needed to make sense of it and put it into context.

Information extraction for semi-structured texts. Fact Book for question answering.

Wiki extraction. [2]

Using clustering and concurrence clustering.

More complex is extracting evidence from many Web pages. Combining evidence from many sources.

Association rules. Market-basket analysis attempts to find interesting relationships among concepts.

Attribute Extraction and Ontology Extraction

It is difficult to build an ontology (2.2.2). Perhaps that could be automated. This would be useful for question answering.

More ambitions is use the templates into a collection of facts so we could call the entire process "fact extraction". This can be useful, for instance, in automated question answering (10.12.0).

Attribute extraction. But, of course, attributes are not always clearly defined for categories (2.1.3).

Ontologies are very costly and time-consuming to develop. It would be helpful to extract them automatically. Find local context such as related words in a sentence.

Argument Extraction and Discourse Processing

Use templates to identify discourse structures (6.3.2). OPV. Discourse markers are often subtle and difficult to extract automatically.

Examples

Figure 10.38: Difficulty of extracting opinions. Irony

Detecting Opinions and Differentiating Conflicting Viewpoints

Based on discourse analysis (6.3.2). Attitudes (4.5.2) and affect (4.6.2).

Detecting buzz (8.4.3). Sentiment analysis. Movie reviews. Restaurant reviews.

The basic level is fairly easy: Mining opinion words and valence words. Using templates and heuristics. This can be used for processing phrases such as "The food is good".

Extracting opinions versus extracting facts. Opinion summarization. Feature-by-feature comparison. Determining the spectrum of opinions.

Considerably more challenging is the extraction of satire, sarcasm, and irony.

Review 1. We stayed here for a weekend trip. We checked in around 1pm but they didn't have any rooms ready so we had to wait about 20 mins. When we did get our room, it was worth the wait. Very spacious and modern design. Clean and comfortable. The fridge and microwave were very handy for us.
Review 2. This is the second time I have stayed at this hotel. I was in room 204. I thought it was average the first time. I was less than impressed this time around. The room looked dirty. Then I start tearing the beds apart to look for bed bugs (thankfully I found none) but I saw a long strand of hair.

Figure 10.39: Reviews example.

10.5.4. Content Analysis

Qualitative and quantitative content analysis. Often a numeric analysis of properties. Sentiment analysis (10.11.2). Discourse analysis (6.3.2).

Word Bursts in Text Streams

[35] Query streams. Communication models.

Resolving Disputed Authorship

Authors impart very distinctive characteristics to their writings. The *Federalist Papers* are essays published in the 1780's encouraging the adoption of the U. S. Constitution. They were written by Alexander Hamilton, John Jay, and James Madison. Most of them were published under their author's name. However, the authorship of some of the papers is a matter of debate; some historians attribute a given work to Hamilton and others to Madison. By analyzing the word frequency, we might identify the author's characteristic pattern. An analysis of word frequency applied Bayes Rule (-A.8.2) to characterize the words selected by the two authors[48]. This Bayesian model was then applied to the essays whose authorship was uncertain and the author was identified with a high level of confidence. [38] This is a type of behavioral signature.

A fifth desideratum illustrating the utility of a Senate is the want of a due sense of national character. Without a select and stable member of the government, the esteem of foreign powers will not only be forfeited by an unenlightened and variable policy, proceeding from the causes already mentioned, but the national councils will not possess that sensibility to the opinion of the world, which is perhaps not less necessary in order to merit than it is to obtain, its respect and confidence.

Figure 10.40: The beginning of Federalist Paper #63 whose authorship was originally disputed. The Bayesian statistical analysis assigned this essay to James Madison.

10.5.5. Literary Analysis

Stylometry.

How do authors affect each other's writing.

10.6. Text Categorization and Summarization

We described indexing earlier (2.5.3). Automated indexing can be considered a language technology.

10.6.1. Text Categorization and Classification

There are many aspects of language which need to be categorized: POS, topics. speech acts (6.3.1). The detection of spam (10.3.2) is a text categorization problem — text filtering techniques are not wholly successful at identifying the subject or type categories of email (3.2.3).

There are two ways in which text categorization can be accomplished: through a priori categories and through ad hoc categories. A priori categories are predetermined categories; using them, a document is placed into the category that best fits its subject matter. This is a general classification system (2.5.1) or subject hierarchy, such as the Dewey Decimal Classification system. Relevancy signatures[56] may be included to illustrate how strong is the thematic link between a document and the category that contains it[??]

Adaptive categories are created in response to the subject matter of different documents. This method may give a more descriptive account of documents content, but it is also difficult to locate such adaptive categorizations into a larger system.

Statistical models for classification. (-A.11.2). Feature extraction. When a set of examples is available, they may be used for training (-A.11.0), to categorize texts without human input. Word distribution, frequency statistics, and overlapping word strings can be combined with information filtering (3.2.3) to facilitate the creation of document hierarchies. Other methods include Bayesian learning algorithms (-A.8.2) and linear regression models[69].

10.6.2. Text Summarization

Summarization has many uses: meeting archives could be summarized for fast comprehension, instructional and how-to books could be summarized to identify particularly apt sections, stories can be summarized to identify interest, etc. It is often helpful to get a quick impression to understand a long text in a short time. Supporting a variety of inter-related information needs. Relationship to abstracts (2.5.5) and tutoring (5.11.3).

Summarization, though seemingly simple, is a very complex process that speaks to the core of language understanding difficulties. Summarizations generally need to be based on semantic relationships and logical order. It is helpful to start by identifying the structure of a text, and likely its summarization. Multimedia summaries. Problems of including different information.

Extractive Summarization

Extractive summarization plus smoothing. Abstractive summarization and text generation. Problem of pronouns.

The simplest approach to creating summaries simply extracts passages. One statistical method for this uses $tf{\cdot}idf$'s, which are term-weighting measures used for text retrieval (10.9.2), and stand for term frequency and inverse document frequency, respectively. Sentences within a text that have the highest $tf{\cdot}idf$ are considered pertinent, and can be used to create a summary.

Query-term methods.

To a limited extent, this can be achieved using concept maps and discourse. Other methods, such as statistical and rule-based systems can also be used. Statistical analysis, using semantic understanding and word charts, can be performed on a text and can, to a limited degree, determine the most useful sections to keep in a summary[59].

A particularly effective technique is to take intermediate frequency words[42] and find sentences with those words (10.9.2). Templates can be helpful in summarization.

Extraction followed by fusion. Among the problems is minimizing redundancy.

Abstractive summarization. Use a semantic grammar (6.2.3).

Task-Specific Summarization

A summary might be constructed as a response to a user's query. These can range from current awareness abstracts (2.5.5) to responding to specific user's query[4] and explanations (6.3.4). This latter is similar to question answering; though, the content is based on a single document. This entails not only the difficulties of automated summarization, but also those of question answering. In a sense, this would function as both a simple search engine, pulling up documents that fit a user's query, but it would advance a step further and summarize the most important information that those documents contained.

A summary could then be made interactive. That is, a user could highlight information within a summary that was either very useful or not useful, and the system would re-generate the summary based on that feedback. As a particular individual or category of individuals (based on job type, for example) use a system more, that system could generate a model of their information needs and retrieve and generate information according to those parameters. This effectively becomes a tutoring system which tends to put answers in a context that a user can understand. Indeed, this type of summarization can be seen as similar to query-oriented question answering (10.12.0).

Summarizing Multiple Text Documents

Beyond coordination of documents to coordination between collections. Certain elements of query-based summarization require that an information system be able to glean information from multiple sources and synthesize it. This could also be termed "document compilation". Multiple document summarization[45] provides a summarization of the information contained in a collection of documents. This is not only helpful for query-based summarization, in which information pertaining to the query may not be contained in a single document, but it also proves helpful to give an indication of the differences among related documents[43].

News summarization often gives extra weight to the ordering of information since in news stories the most important information is often presented early. Topic themes in multi-document summarization. Discourse processing.

Comparing articles.

10.7. Search Engine Interfaces and Interaction

Search and information needs (3.2.1). Search engines typically support simple queries. Document structure (2.3.1). Taxonomic organization (2.2.2). In the past few years, search engines have gone from being a relatively obscure academic specialty to the transformation of society. Cognition and usability (4.8.0). HCIR. Combining social search with search engines.

Search results presented via an app.

Conceptual models for interaction with search engines [**?**].

10.7.1. Using Search to Interact with Collections of Text Documents

Searching is one way for a user to interact with a collection of documents. Indeed, the Web would barely be useful without search engines. In these sections, we are primarily concerned with how this can be enhanced. Indeed, this interaction might be seen as a conversation; a user asks a question of a collection in the form of a query, and the collection respond based both on its content and the manner in which the question was phrased. A searching interface should be designed to support all aspects of this conversation. Because information needs are often complex search interaction can be very complex. Beyond the desktop metaphor.

Anchoring a query with one term and then making revisions to it.

Understanding how a person searches by following eye movements (Fig. 10.41).

The full range of interaction involves a combination of the user, the task, the corpus, the interface,

Figure 10.41: Eye movements during examination of a search engine results page. (from Cornell). Note that the user is focused almost exclusively on the first two returned results. (check permission)

and the retrieval algorithms; having access to this complete range of aspects allows the user to address larger tasks and to build information environments, and not simply to search (3.5.4). The ease of a user's interaction with any of these aspects is largely determined by interface design.

Privacy and browsing history.

Search Engine Query Logs

Can be used for targeted advertising or for improving services.

10.7.2. Using, Forming, and Modifying Search Engine Queries

Earlier, we considered strategies a trained reference librarian might use to improve queries. Based on information needs (3.2.1). Here, we can consider query management in terms of ranked retrieval. A large number of queries consist of only one word[14]. This can be difficult if my query were "Russia" or "Beatles" what would you infer as my information need? What would be the results to return. A particularly wide range on issues for Web queries (10.10.2).

Earlier, we considered systematic search strategies (3.3.1).

Query characteristics. One or two words.

Interface tools can help a user develop effective queries. The difficulty users have when constructing Boolean queries can often be alleviated by utilizing graphical interfaces. These could be designed in a way to first support the creation of a query and then aid in its reformulation for more pertinent results.

Query Categorization and Analysis

Query signals: "Swiss Baker Alpena" For instance, names are most often entered in natural order. Query term order.

Like question categorization (2.1.1), it is also helpful to categorize queries. Automatic processing of the query. Does this query include a proper name? This is analogous to the discussion of question types (3.2.3).

Query Expansion

If we have a query with the term "car" it is probably reasonable to expand that query with the term "automobile" so that both terms can be presented to the search engine. There are often difficulties with exact-term techniques because alternate semantic forms of the words are not included in the search. Or, there are other plausible search terms. One approach is to use thesauri (2.2.2) to expand the query. The query term "boat" might be expanded with "yacht," "barge," "freighter," and "ship". This would increase the number of documents returned, while at the same time preserving their pertinence. Unfortunately, it is also possible that the wrong word sense of the query term will be selected, and the query expansion will generate irrelevant words. This often occurs with homographs, such as wind (to wrap around) and wind (breeze).

Example of query expansion. Fig. 10.42

Query expansion.

Figure 10.42: Examples of query expansion.

Relevance Feedback

Relevance feedback employs terms from documents which the user has indicted as the most relevant one retrieval. Frequent terms from those documents could be added to the original query. after one set of responses has been received from the search engine, the user can select what items have the most relevance, and the system will use these rankings to help the user construct a more accurate query. A user asks the system to return "more documents like this" (Fig. 10.43).

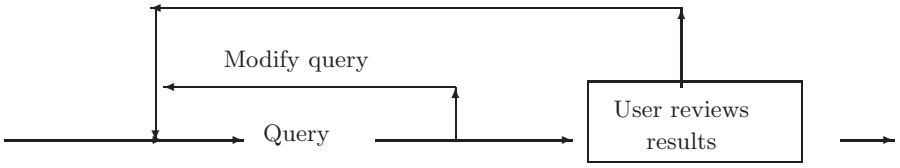

Figure 10.43: Relevance feedback example.

In addition to positive relevance, negative relevance feedback must also be included when using relevance feedback to modify a query; the characteristics of documents selected as "not relevant" can be used to restrict the type of documents returned on subsequent retrievals.

Query Development and Re-formulation

Many routine queries are reformulated. Query assistance.

As we noted when first discussing search, users will frequently revise their queries. Fig. 10.44 illustrates how a manual document query reformulation might occur. Other difficulties with queries can be more easily solved; spelling errors in queries can be fixed by the inclusion of a dictionary program into the text search field.

Query development and controlled vocabularies.

Interfaces,
Stages of reformulating a query.

Figure 10.44: Examples of reformulating a query.

Supporting Complex Queries

Complex queries (3.2.3). Using search histories.

10.7.3. Examining Search Results

Following a search, the user browses. Search followed by browsing. Combine aspects of the two processes with a graphical interface. These are tools which could be attached to an information workspace or a document management system.

Terms in context. Identifying results and telling us more about them.

This gets us into the more complex task of retrieval of highly complex searches (3.2.2).

Document Surrogates

When the documents are returned, they need to be described by surrogates so the searcher can judge whether the documents with which they are associated are relevant to the information need (3.3.3). Typically surrogates include metadata and brief metadata descriptions or summaries (2.5.5).

Visualization Search Retrieval Interfaces

Although ranked retrieval is simpler, some sophisticated users prefer Boolean retrieval procedures, because it is easier to understand how to modify the search, as this is part of the users' models for how the search engine works. We will see general visualization interfaces (11.2.5).

Duplicate detection for web page search engine output.

These interfaces do more than simply showing surrogates, they can provide interactivity for rapid exploration of the result sets. In particular for comparing different attributes.

The Vibe[52] interface for browsing the return document sets (Fig. **??**). There are reference or anchor points among which retrieved documents are positioned. In this example, a set of eleven documents is positioned across a space generated by three terms.

May perform set operations on query returns.

Viewing the distribution of terms within a document.

Categorizing Search Results

Categorizing search results is essentially developing a table of contents for the results. Automatic classification. Part of categorization involves clustering document return sets[12] (Fig. **??**).

Similarity within categories.

It helps to have an established metadata set.

Sometimes there is a focus on individual aspects of searching, sometimes on entire tasks. Coordination across windows. As noted earlier, managing the search may be contrasted with managing the task completion (3.0.0). The focus of interface design can move from supporting word search to full task support. Desktops can be maintained with queries in progress.

10.7.4. User Issues and Interfaces for Web Searching

Search engines have proven to be very powerful but they do not support complex search very effectively [30]. Because the Web is so varied, perceived effectiveness is often important for commercial search engines, which depend on their customers believing they have done a good job. Search behavior (3.2.3). Re-visitation.

Applications of searching in a heterogeneous knowledge sources such as the Web. Web site design Earlier we described the Web as a common-use hypertext, (2.6.3).

Web Page Filtering. Some Web pages, may be dropped from search engine results. This is similar to spam filtering (10.3.2).

Interfaces for Web Searching

Unlike collections of separate documents, Web documents are richly linked together and it should be useful to maintain that linking in access. Just as we considered interfaces for browsing the results set of a text retrieval interface, we can consider interfaces for browsing Web searches. These are not simply hierarchical connections.

Providing context for search hits.

Visualization tools for showing search results.

Related Web pages which are retrieved may be grouped together even beyond ranked retrieval[18]. It presents a hierarchical view of pages that reflect the internal structure of Web sites that have pages that match query terms.

10.7.5. Web-Based Collections
Using a search engine to find child-friendly materials.

10.7.6. Personalizing Search
People often repeat searches. Problem with personal relevance.

Desktop Searching
Desktop searching. Such as Gmail. Personal information management (4.11.0).

10.8. Web Search Engine Business Models and Policies
Because searches often reflect a searcher's needs, a search engine is a good place to advertise. This can be done by auctioning search terms. The search engine company needs to guarantee neutral rankings (8.12.5).

Search result ordering as free speech.

Sponsored search. Advertising campaign. (8.12.5). SEO (8.12.5). Because many users are led to Web pages by search engines, some Web site designers who want their sites ranked by those search engines add spurious text to Web pages that will be picked up in Web indexing processes. This is known as "keyword spamming". A variety of techniques have been developed. These include link farms.

Search results as free speech or as a utility. Neutrality in search engine results. Avoiding bias. Transparent search ranking policies.

Search engines presenting service directly rather than search results.

10.9. Search Algorithms

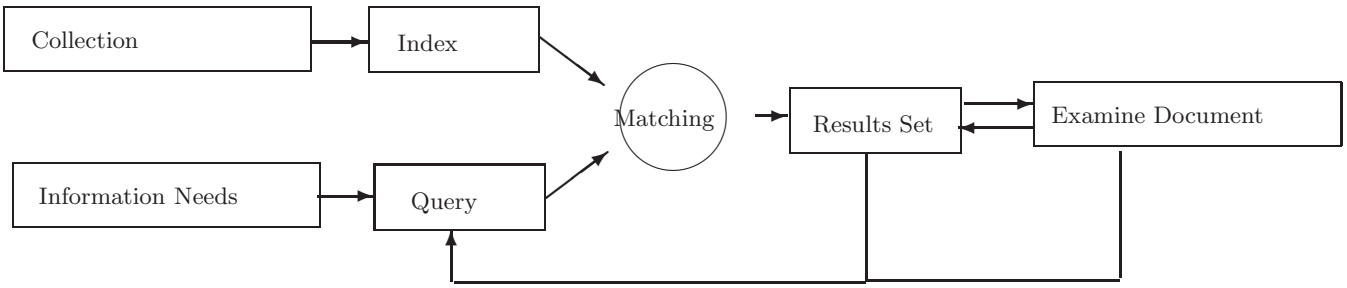

Figure 10.45: Recall that during search the user's needs have to be coordinated with the indexing system.

Represent a document in the context of a collection.

Term and vector indexing. Social media. Linked data.

Standard search methods tend to employ exact matching, which returns sets of documents based on exact keyword matches. Statistical methods, on the other hand, use various formulas to match documents to a query based on statistical similarity. Perhaps surprisingly, the techniques of natural language understanding and processing generally (10.4.2) are not as effective as purely statistical techniques. The decision to implement a text retrieval system is not based only on effectiveness. Many factors, such as effort and cost (3.3.3) can guide system selection.

Normalization and indexing.

10.9.1. Content-Based Models: Text Retrieval via Term-Matching
The simplest techniques approaches treat text retrieval as a data retrieval problem. Significant words from a text are stored in a database, and documents containing those words are returned in response

to a query. To return all the news stories about "California," exact term-matching would find all news stories that contained the word "California".

The inclusion of metadata searching can increase a searcher's effectiveness. A search that is limited to the words that appear in the text often have difficulties returning pertinent results. One difficulty is the return of spurious matches. Using the example above, spurious returns might include stories that referenced California Avenue in Washington D.C., or California University, which, curiously, is in Pennsylvania. Alternatively, when multiple terms are in the query as keywords, we say the query is "over-specified". While there may be a chance that all of the selected terms will be present in a single document, it is much more likely that nothing will match the terms exactly, and no documents will be retrieved. This approach is similar to database retrieval although it may be made more efficient with data structures such as an inverted indices.

The basic term-match model can be extended in several ways, such as utilizing Boolean (3.9.2) combinations of terms. Other ways of applying the term-match model include adjacency and proximity operators: "immediately adjacent to," "within the same sentence as," "within three words of," etc. An illustration of the usefulness of these types of searches is that a search for the Federal Bureau of Investigation is quite different from, and will return very different results, than a search for a "bureau of federal investigation" or the "investigation of a federal bureau". This can be a phrase search.

Simple term-matching can be extended in several ways. These may search various fields of a database or a collection's metadata and allow users to specify search criteria for each one. The more search criteria that are selected, however, the greater are the chances of over specifying the search and retrieving no documents; these are known as fielded searches.

Proximity search.

10.9.2. Ranked Retrieval and Term Weighting with the Vector Space Model

Both simple Term Matching and Boolean techniques return an unordered set of documents — any document that contains a term that corresponds to the search criteria will be retrieved, and a document that contains one instance of the term will be given the same weight as a document that contains five instances of the term. Ranked retrieval methods attempt to rank documents according to how well the system believes they would answer the user's question. Generally, these ratings are based on a statistical measure of similarity.

Representing the Corpus with a Term-by-Document Matrix

As we have seen elsewhere, the representation is the key for information systems. For the task of searching documents, the representation must capture the richness of a complex document in a way that allows matching relatively unpredictable queries. Essentially, these are distributed representations for the contents of a document or book. Problem with orthogonality assumption.

Zone indexing.

The Vector Space Model is the t-known procedure for ranked retrieval[60]. With it, both the query and documents are represented as vectors. This is also known as a "bag of words," since the order of the words is discarded. This algorithm does not preserve our intuitions about natural language very well, but it is the most robust technique.

Taken together, the set of vectors from the documents in a collection form a "term-by-document matrix" Fig. 10.46 shows a hypothetical term-by-document matrix for a document collection — such matrices for actual collections may include thousands of documents and terms. This is a representation of the collection (1.1.2).

Matching Queries to Documents

The details of the calculations for the Vector Space Model are given in (-A.6.3). Suppose that we wanted to find documents that matched a query "automobile tires". When there are several terms in the query

	Document						Query
Term	D_1	D_2	D_3	D_4	D_5	D_6	Q_1
boat	1	2	0	0	1	0	0
boats	3	0	7	0	0	0	0
sailing	4	1	1	0	1	0	0
water	2	5	3	0	0	0	0
car	0	1	0	0	6	2	0
automobile	0	0	0	4	0	5	0
truck	1	0	0	1	3	0	1
tires	0	0	0	4	0	2	1

Figure 10.46: The term-by-document matrix is one way of representing the documents in a collection. Each document is a vector and the query "automobile tires" is also represented as a vector. The query vector is compared to each of the documents for the match. The match would be D_4.

and different frequencies for each of those terms in the document, it is necessary to determine how to weight the terms in the document to best match the terms in the query. Two particularly effective weight-determining equations are the term-frequency (tf) and the inverse document frequency (df). Eq. 10.2 shows tf and is illustrated in Fig. 10.47. Fig. 10.48 shows df. Eq. 10.3. These actually are calculated as $\frac{tf}{df}$ the inverse of document frequency, $\frac{1}{df} = idf$ is usually written as $tf{\cdot}idf$. A complete calculation of $tf{\cdot}idf$ is given in -A.6.3.

$$tf_{simple} = \frac{number\ of\ times\ the\ term\ appears\ in\ the\ document}{total\ number\ of\ terms\ in\ the\ document} = \frac{t_d}{T_d} \tag{10.2}$$

Figure 10.47: A document should be more likely to be returned in response to query terms if contains several occurrences of a query term (right) than if it contains only one (left).

$$df_{simple} = \frac{number\ of\ documents\ with\ the\ term}{number\ of\ documents\ in\ the\ collection} = \frac{D_t}{D} \tag{10.3}$$

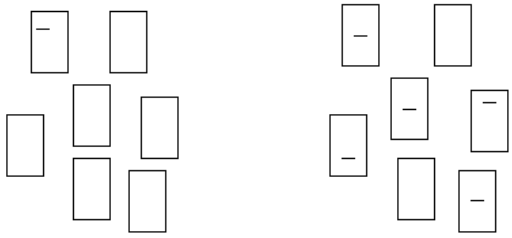

Figure 10.48: A document should be more likely to be retrieved in response to a query term if that query term (indicated by a short dash) occurs in only a small number of documents (left) than if it appears in many documents (right).

There are many extensions and variations of the Vector Space Model ((sec:additionaltextretrieval)). In addition, techniques related to the Vector Space Model are useful for retrieval of other types of content such as color matching in images. matching notes in music ((sec:audiovectors)), and for collection selection. Precision and recall (3.3.3) can also be used as metrics for the effectiveness of an indexing algorithm – such as the vector space model – to match a user's queries

These measures can also evaluate the performance of a similarity-based ranked retrieval engine. They are usually inversely related: as more documents are retrieved, fewer of them are likely to be relevant (Fig. 10.49). People are not very good judges of the number of relevant documents in a collection and the performance of a related system in retrieving them[13].

Total documents in collection	100	100	100
Number relevant in collection	20	20	20
Number retrieved	10	30	60
Number relevant and retrieved	5	10	15
Precision	0.50	0.33	0.25
Recall	0.25	0.50	0.75

Figure 10.49: Example calculations for precision and recall when retrieving 10, 30, and 60 documents. Usually as more documents are retrieved, the precision decreases and the recall increases.

10.9.3. Statistical Representations Beyond the Vector Space Model

Retrieval Based on Probability rather than Similarity

Probability rather than similarity. The most effective models for search engines are statistical. Because models such as the Vector Space Model are based on the statistics of words in documents, it is worth considering the properties of language. As might be expected, the word "the" is the most common word in the English language. Conducting a search for that word matches too many documents.

The *idf* value is another way to describe the frequencies of terms in a set of documents. While we've shown that the most common words in a language can be useful for text retrieval, the *idf* value uses the least common words for the same purpose. People generally express themselves using only relatively small set of words, and these words tend to be similar from person to person. Therefore, these words are used in the context of a wide range of different types of documents, and are thus not very specific. The *idf* value uses the *least* common words to identify relevant documents under the idea that the less common they are, the more specific they will be.

Zipf's Law (-A.10.2). For instance, this can optimize index compression. Derived from least effort [**?**].

Ideally, retrieval systems would be based upon a model of the properties of natural language. Most of these models, however, would depend on the structure of the training corpus. Different disciplines use different writing and communication styles and, to an extent, a different vocabulary. Accurate document retrieval is thus based on language models of the type being used in retrieval that accurately established the frequencies of word occurrences and co-occurrences. This requires a large and specific training corpus for the results. The text-processing principles used for most systems are derived from statistical measures of the language of these corpora.

Distributional semantics.

Various analyses have shown that words in the middle frequencies are the most predictive and useful for text retrieval. Furthermore, within these useful words, two different kinds of words may be identified: "function words" and "content words" (Fig. **??**). These roughly track to the different functions that verbs and nouns play in a sentence. Poisson-distribution.

Another method for utilizing statistical information for text retrieval generates statistical concept spaces. These define the semantic relationships between concepts for a given corpus. Then, these relationships are used in conjunction with standard statistical analyses to produce more accurate retrieval results.

Language Models

Topic Models
LDA.

10.9.4. Search Based on Machine Learning
Many queries are regularly repeated. Search engines from massive data sets and machine learning rather than from the traditional analysis of text. So many searches are repeated and it's possible to collect large numbers of searches. How many people are asking the same question.

Algorithms for learning. Machine learning (-A.11.0). Spell corrections (10.4.1).

Rank position.

SVM and learning similarity.

Humans in the loop for high-frequency searches.

10.9.5. Indexing and Searching the Web
The Web is a particularly chaotic environment compared with organized document collections. The most common application of text retrieval is Web search engines. The Web seems to encourage keeping multiple versions of a document on different servers. Here we focus on text search engines, but many of the principles also apply to multimedia Web search engines. Manage several components: index, retrieval. The Web is a common-use hypertext environment (2.6.3).

Semantic information added explicitly to web pages. Schema.org and microdata elements for indexing the web.

10.10. Characterizing and Indexing the Web
10.10.1. Dynamic Content on the Web
Some Web pages are frequently updated. Some pages change a lot but others do not. Simil Staying on a page. Words on a page that are unique to a page. Curated Web pages and analogous issues addressed in the archival literature (**??**). Changes within pages. Indeed, some types of pages are expected to change. Fig. 10.50. Expected changes versus unexpected changes. For instance, breaking news (8.13.7). Preservation of changing web sites.

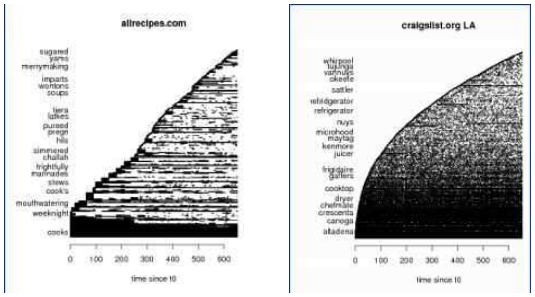

Figure 10.50: Two examples of term-longevity graphs (from[21]). (check permission)

10.10.2. Crawling and Indexing the Web
Because the Web has no systematic overall organization, search engines often are the most practical way to find pages. The machine index of Web pages used by a search engine is usually created by a Web "robot" or "crawler," which follows engine links to visit as many pages as possible. Pick a page and explore all links from there. Dynamic web (10.10.1).

Web characterization (2.6.3).

The Web is frequently updated. "Link rot" occurs when a Web page changes its location so it no longer links to it are no longer accurate.

A second problem is that some content is frequently repeated on many Web sites. Redundancy.

Targeted Web search. Focus on those pages which are most promising.

The software that generates these pages is known as a "robot" and the process is sometimes called "spidering" ar ?crawling?. Many pages are not indexed — the robots cannot get by passwords or image maps. Furthermore, Web pages can also be marked "robot exclusion" meta tags to request that the pages not be indexed.

<META NAME="ROBOTS" CONTENT="NOINDEX, NOFOLLOW">

Crawling policy. There are many databases whose contents cannot be searched by Web robots. This may be because of security or simply because the Web pages are generated dynamically. This content is called the "deep Web". In some cases, appropriate responses to the database query can be inferred from the page.

Duplicate detection.

Measuring Web Traffic

Most visited sites. This is an issue for search advertising.

Characteristics of Web Queries

Because the Web is so broad, and it has so many different resources, Web queries are very broad. (3.2.3). e to explain why a specialized database would

Search tools when there are ambiguous query terms.

One strategy for processing queries would be to categorize them. For instance, whether a user is searching for a document or an Internet service. Web queries are often quite broad. 1) apple, peach, kiwi 2) Moscow. How might you categorize each of these queries? What would be the optimal response by an information system? A user may be trying to find not just a document, but a service as well.

Perhaps not surprisingly, the most frequent searches deal with entertainment and sex.

Types of Web searches[24][58] with different information needs.

Using Web Page Links as an Indication of Similarity

Web communities are powerful predictors of association (2.6.3). The Web has aspects of both collections and hypertext. With most search algorithms, only the text of the document itself contributes to the likelihood of a document being retrieved. We identify many types of links between Web pages can indicate the similarity among them. Effectively, this is a principle of social retrieval. This is comparable to citation analysis (9.1.2). If the central node has links from other important nodes, it should acquire a high value (Fig. 10.51). The PageRank algorithm [**?**] provides a mathematical solution for this problem (-A.6.5). This type of analysis is also useful for building "family-friendly" indexes since adult-only sites often link to each other and these can be filtered out.

Relationship of journal impact (9.1.3) to PageRank.

In the context of a document structure. Path queries.

Using social networks to find associations (5.1.0).

Indexing Large Collections

Because the Web has so many documents and because there are so many hits on search engines, search procedures must be very efficient. Indeed, several shortcuts may be taken. The retrieval algorithms are not necessarily used. Web archives (7.5.5). Distributed data centers ((sec:datacenters)).

Web pages change frequently which if different from typical documents.

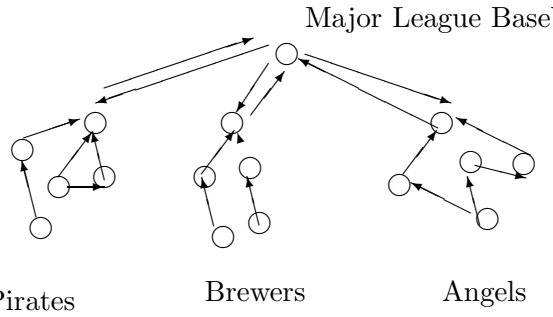

Figure 10.51: The PageRank algorithm includes the link structure in determining search hits. The Major-League Baseball site might be a good match for a query on "baseball" because it is centrally located among many other sites such as the names of individual teams (Pirates, Brewers, Angels). (redraw)

Furthermore, the results should be matched with the needs of real users. Many queries on Web servers are repetitive. The same user may simply repeat a query, or several users may be interested in the same thing. For common queries, the results may be saved in a cache so they can be retrieved later; these responses can even be hand-tuned (-A.14.2).

Duplicate detection.

Efficiency of building an index for very large collections. Often across several machines.

Real-Time Indexing

Indexing Tweets.

Tweet examples.

Figure 10.52: Tweets

Updating search engine results. Global data centers (7.7.3).

10.11. Beyond Basic Search Algorithms

Text retrieval has proven so useful that we can extend it in a broad range of content types. Social networks (5.1.0). Recommendation systems (5.5.5). Knowing more about searcher's context. Semantic search.

User Context and Search

Using location and context to determine searcher intent.

10.11.1. Social Search

Merging searches, and social media.

Use of context based on friend's interests. (5.1.4).

For what items do you value your friend's opinions. This shares some aspects of a recommendation system ((sec:recommendation)).

Finding People and Extracting Information

Representations of people. Social perception (5.5.2). Finding people with similar interests. Blogs, Twitter, Facebook. Dating. This is a slightly difference sense of social search than we encounter before (3.3.2).

Finding People with Expertise

Sometimes when we cannot find the answer to a question, we can find a person who might be able to answer that question. People are very helpful at providing information. People finders help to locate people with specific attributes (5.1.3). Knowledge management and organization charts. One simple

approach could be to enumerate all the attributes and attempt to describe these simple database listings with a controlled vocabulary. True expertise versus claims of expertise.

Representing expertise. Expertise for different topics (3.7.1). Multiple dimensions of expertise. You may need to find somebody who speaks Urdu or somebody who knows how to build a house. Expertise network as a social network (-A.3.5) or community of practice (5.8.2).

You could search for your human expert in much the same way you would search for a document[36]. As with other information search tasks, representation is an issue here. The expert needs to be identifiable in the same way a document might be; that is, the expert's credentials would be represented in text and retrieved by a search.

Privacy issues in people search.

Information "logistics". Getting critical information to an analyst.

Organization Roles and Structure

As we have observed earlier. Consider all the information artifacts that an organization generates. Could we sift this to for a view of the organization. This is an aspect of enterprise content management (7.3.6) and text mining.

Intranet searching.

Personal information services (4.11.2). Social filtering (5.5.5). Email tasks (10.3.2). Email can be very messy.

Beyond simple documents to heterogeneous sets of materials.

Finding people's home pages.

Privacy concerns.

Summarizing email.

Fig. 10.53 shows empirically derived social network with a pathfinder network.

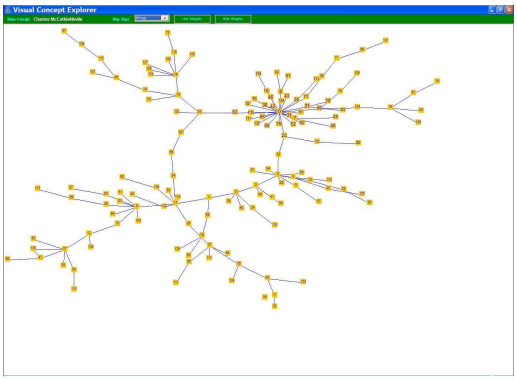

Figure 10.53: Social network derived from a set of email messages[71].

Argumentation pro and con.

Getting away from the traditional view of documents.

Interaction in virtual and loosely structured organizations (5.7.3).

Using evidence about what kind of searches other people are doing and what sort of selections they making from those searches.

10.11.2. Processing News and Other Content Streams

A lot of content evolves across time. Text streams. real time.

News, enterprise search, affective materials.

How should time be weighted as a factor in interactive information systems?

Lipstick on a pig (5.1.3).

Social curation and media aggregators which select content from media streams such as blogs.

Processing News

News (8.13.7). Provides a challenge for text processing. Newsblaster and summarization (10.6.2). Temporal summaries.

Change detection.

Citation analysis which we discussed earlier (9.1.2) can also be thought of as looking at streaming content.

News is an almost continuous stream of information. Imagine trying to track a news story from day to day in a newspaper. There are often new wrinkles and even stories that split off to form other news stories. Some question whether a "topic" is a distinct entity. We would like to group events to identify the story. It is basically a cluster analysis to show whether a document falls into an old cluster or a new one.

Content streams are particularly challenging because the content changes frequently. This makes it difficult to calculate the $tf*idf$ formula. News may develop on unpredictable topics from many sources, so topic detection is important for broadcast news services (10.11.2).

Judging relevance of news stories adds the dimension of recency to the user's interest.

News versus newspapers. Newspapers have more than formal news. In fact, a newspaper has a complex collection of content such as classified advertising, sports scores, and weather reports.

Video news.

Ranking these might include weighting the articles by recency. Articles for newspaper will be clustered.

Sentiment and Buzz Analysis

Public opinion (8.4.3). Because the Web is now so interactive, news items and trends are widely discussed online. Check social media sites.

Information sharing in virtual environments.

Blogs.

Blog Quote

Text analysis of blogs. Determining what makes a positive review of negative review finding affective tone – instance in music reviews. Epidemics and the spread of information. This is a type of information diffusion.

Many domains: from product reviews.

Counting polarity. Polarity classification. Difficulty of using single words. "I think that is a spectacular idea..."

Many blogs are full of grandstanding and gamesmanship and this has to be unwound before any valid analysis can be conducted. True identify is often unknown (5.5.1). As a community norm, blog contributors can be anonymous except when then have a vested interest in the claims made in the blog. The latter are known as sock-puppets. Blogs anonymous authors. Blogs and information overload.

Cultural and population effects. (Fig. 10.54)[70]. Google Trends. Correlated with news.

Using search terms to make predictions. Even to the point of basing stock market trades on sentiment [41].

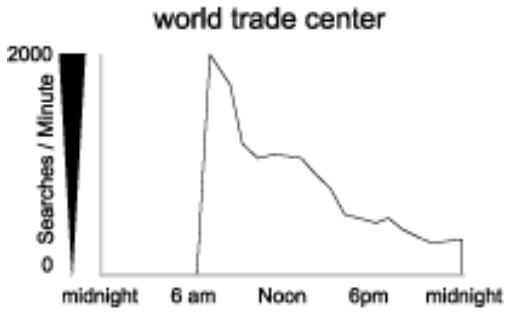

Figure 10.54: Searches on google.com from the "World Trade Center" during September 11, 2001[70]. Note that the attack there was just after 9AM ET (6AM PT). (check permission)

Buzz analytics. Opinion words (10.5.3). Sentiment flow. Sentiment combined with discussion of aspects.

Figure 10.55: Web ratings.

Monitoring the effects of advertising with search engines.

Detecting agreement in conversational. (6.4.0).

Detecting Misinformation and Deception in Language

Deception (5.3.3). Deception in voice.

10.11.3. RDF Search

Searching XML triples and RDF. Mixing RDF and semantic web, Incorporating structure into search.

10.11.4. Personalized and Social Search

Personalized search is a special case of social search. Representation of personal interests. Personalized selection of news articles. Many under-specified search terms.

Combine recommender systems and social search with page rank to make predictions for every individual on the web about their search topics.

10.11.5. Specialized Search

Search for specific categories of items such as search for personal names. We expect the order to be important.

Search semi-structured documents. Patent search. Family search.

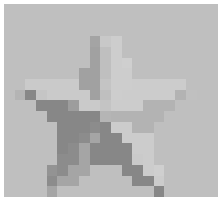

Figure 10.56: Family search.

Search engines which return responses fitting a certain religious or cultural sensibility.

10.12. Automated Question Answering

Questions are the natural way we use language in social interaction to get information. Questions are different from queries, in that a query is processed by selecting the words in it, and a question is answered based on the totality of the question's meaning. Document management. Information needs are often formulated as questions (3.2.1). Question answering is a step toward more complex interaction, such as verbal dialog systems (11.10.4) and tutoring systems (5.11.3). Answering complex questions often leads to explanations (6.3.4).

Question answering by reference services (3.3.2). Determining what the user really wants (3.2.1). What makes a good answer[39].

Question Categorization. As we discussed for reference interviews, categorizing questions is a useful first step in answering them. Earlier, we had described a category system used for reference interviews (3.3.0).

For factual questions, the most obvious category scheme is the familiar journalistic questions. Simple "who," "what," "where," "when," and "how" questions are fairly stylized and can, sometimes, be processed automatically.

This generally occurs by generating a query from a question. The question "Who is the Queen of Sweden?" might generate the key words "who," "queen," and "Sweden". From that, the system would have a formalized, yet accurate, accounting of what the question sought. Another simple approach, only functional in particular information areas, is to pre-store answers to common questions.

Using Information Extraction for Generating Answers. Creating an index with information extraction. Extract responses from fact networks (2.2.2, 10.5.0). A quality information resource is needed.

Closed domain versus open domain. Domain-specific question answering questions. Easier for specific questions.

Fact-based questions are the only practical type which can be answered.

Source of the data with which to answer the questions. Mining information from the Web.

A question answering system can use parsing to determine question components which could then be processed with by decomposing the structure of the query (3.3.1). Such a system's first step would be to analyze the question and identify the type of query with simple natural language processing, which involves grammar parsing and ontologies (2.2.2). The system would then need to search for the appropriate information in existing documents or other sources and pull that information from the sources using information extraction and data mining; Matching questions to questions that have already been answered by human beings.

These techniques are discussed in the following section. Then the information would need to be presented back to the questioner in an understandable manner. Information presentation is a complicated task; a system must be able to determine what information is pertinent to the questions, how to

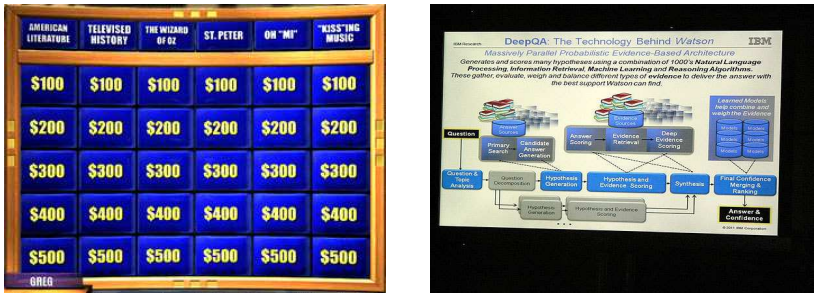

Figure 10.57: Watson is a system from IBM [?] which plays the game of Jeopardy. It is good at handling the ambiguities of natural language but is still does not approach human beings for unconstrained natural language interaction. (recapture)(check permission)

"Married President Washington".

Figure 10.58: Finding answers with approximate queries followed by information extraction.

paraphrase complicated information, and how to normalize concepts so that they are not repeated throughout the presentation. Each of these requirements presents unique challenges.

Automated answering systems should eventually move beyond simply retrieving documents to synthesizing actual answers to questions. This is difficult, as it requires more natural language processing. As we noted earlier, a question answering system may involve the presentation of background information; indeed, it can begin to approximate a tutoring system (5.11.3) and it can organize that information in a way that is comprehensible for the end-user.

Query splitting[5]. Transforming questions to declarative form and then submit to a search engine.

Result aggregation.

Information fusion from several documents. The job of the QA system is to generate

a focused natural language response and not a document.

10.13. Translation and Cross-Language Processing

Language barriers are some of the greatest difficulties in human communication. Translation combines aspects of language recognition and generation. There are lots of differences between languages. Processing different alphabets possess some challenges but that is only the beginning. Transliteration is expressing one language using the alphabet of another.

Voice translation followed by synthesis in your voice in a second language.

Foreign language reading and writing aid.

Language segmentation such as Chinese. Culture differences (5.9.1) are not always readily explainable so it is more difficult to handle and cross-cultural processing.

10.13.1. Translation

As with other text processing tasks, there is a gradation from light-weight methods to rich-text processing (4.3.4). This simplest strategy would be to translate one word at a time. However, this word-by-word translation would be very poor quality. The difficulty of word-by-word translation. More difficult than the translation of terms of linguistic structure is the translation of cultural sensibilities (5.8.2). For example "Step up to the plate". Cultural interpretation. Crowdsourcing translation.

Translating idiomatic phrases can cause a great deal of confusion. The English phrase "art theft" translates into French as the idiomatic phrase "xx"; when re-translated into English, it becomes "flight

Figure 10.59: The Rosetta Stone provided the key for decoding the meaning of Egyptian hieroglyphics because it used three different scripts (hieroglyphic, demotic, and Greek) to display the same text in two different languages (Egyptian and Greek), thus providing a reference point between the tongues. (check permission)

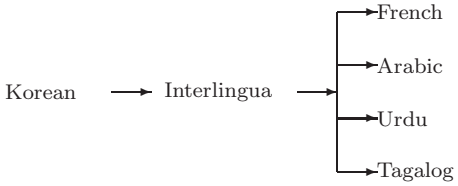

Figure 10.60: An interlingua is a common semantic representation that is independent of languages. It could be useful for translation between languages.

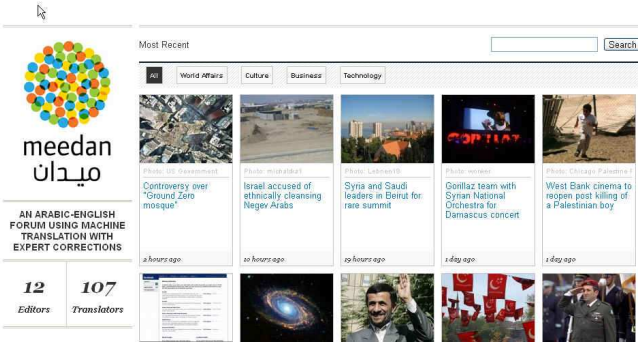

Figure 10.61: Meedan Arabic-English (human) translation blog. (check permission)

of art".

Translation from one language to another stretches the capabilities of language processing. Domain-specific translation is easier than general purpose translation. Domain customization. Tailoring to many languages.

Translation of poetry. Difficulty of finding rhyming. Even aside from poetry, a lot of language is metaphor and metaphor is particularly difficult to translate.

Multilingual chat service.

Interlingua

As with other types of language processing simple methods are often adequate for translation. The simplest approach is surface translation. A lexicon can be helpful in connecting concepts. We do not want to just translate a word, we want to get at the correct sense of that word. A word many translate into a phrase or vice versa; one should aim at an alignment between texts in different languages. An

"interlingua" is a common underlying semantic representation which can then be used to generate text in other languages (6.2.3).

There are different qualities of translations. A translation may simply provide the gist of a passage or the details of a weather forecast, while the translation of a poem may itself attempt to be poetic. Several dimensions for measuring the quality of a translation are shown in Fig. 10.62. This can be used as a criterion for improving the performance of an automatic translation system.

Factor	Description/Example
Adequacy	Are the semantics conveyed?
Grammar	Is it grammatical?
Idioms	Are idioms used correctly?
Fluency	Is it easy to read (speak)?
Style	Is it elegant?

Figure 10.62: Criteria for effective translation[53].

Text-to-text. Speech-to-text. Speech-to-speech.

Multilingual blogging. Multilingual chat.

Machine Learning for Translation

Parallel corpora. Machine learning (-A.11.0).

Wh-questions (i.e., what, where, when, why, how). Fact extraction (10.5.3). Word alignment. Probabilities of word senses.

Get some reference translations and use them to calculate simple statistics. We could use precision/recall (3.3.3) as overlap measure. Long, continuous word strings are good.

Use probabilities of different word-sense translations.

Alignment of parallel corpora.

10.13.2. Cross-Language Text Document Retrieval

Sometimes, we may want to search a set of documents in a language with which we are not familiar. This is another variation of information retrieval. Sometimes the term is not available. Some tasks are multi-lingual; searchers may want to know of the existence of relevant documents in foreign languages. Thus, searches may cross languages (Fig. 10.63).

Figure 10.63: Two models for cross-language document retrieval. Translating the query or the document. It may be better to do both and combine the results from the two approaches.

Suppose an Arabic-speaking searcher wanted to pose a query to an archive of English-language documents. There are two general strategies: the query can be translated from Arabic to English, or the corpus can be translated from English to French. Clearly, it is easier to translate the query than to translate the entire corpus, but because the queries are generally short, a poor translation could have a serious negative effect on retrieval. A unified representation of the corpus (e.g., LSI) is needed, as is normalization of language. It may be most direct to translate between languages and then do retrieval.

As we have noted earlier such categorization in integral to a culture's identity. Traditional monolingual retrieval provides a benchmark against which cross-language retrieval can be measured. It is worth noting, that in some cases, the concepts themselves are different in different languages (5.8.2). A cross-language thesaurus link attempts to link concepts across languages (Fig. 10.64). As with other thesauri (2.2.2), this can be useful for term expansion. Apply query expansion (10.7.2) for both queries and documents. In addition to simple cross-language issues, cross-cultural factors must be taken into account. Thesauri can be useful for comparing similarity or conceptual structures across languages. Apply cross language frame semantics. Cross-language FrameNet (6.2.3).

English	German
simian	Affe
monkey	-
ape	Menschenaffe
timepiece	Uhr
clock	-
wall clock	Wanduhr
standing clock	Standuhr
tower clock	Turmuhr
watch	-
pocket watch	Taschenuhr
wrist watch	Armbanduhr

Figure 10.64: Some examples from a German-English cross-language thesaurus. For some English terms such as "monkey" there is no analogous term in German. In other cases (e.g., "clock" and "watch") there are no single-word translations from English to German; rather, adjective phrases are needed to translate them (adapted from[64]). (check permission)

Exercises

Short Definitions:

Adjacency operator
Alerting service
Click-through
Cognitive organizer
Copy-detection
Deep Web
Dialog
Document surrogate
Emoticon
Fielded search
Grapheme
Hubs and authorities
Ideogram
Image extraction
Intranet
Information extraction
Lexicon
MathML
Meta-search
Orthography
Parsing
Path queries
People finders
Plagiarism
Primary source
Proximity operator (search)
Question answering
Ranked retrieval
Relevance feedback
Streaming content
Term-by-document matrix
Vector space model
Word sense disambiguation

Review Questions:

1. Describe the writing system for (a) Arabic, (b) Farsi, and (c) Vietnamese. (10.1.1)
2. Is this sentence written in a serif or a sans-serif font? (10.1.1)
3. Give the ASCII and Unicode values for the word "media". (10.1.1)
4. List some skills required to be a proficient reader. (10.2.0)
5. Give some examples of "shallow" and "deep" methods for natural language processing. (10.4.0)
6. Calculate the edit distance between DIMENSION and DINEMSION. (10.4.1)
7. Describe the usefulness of ranked text document retrieval methods. (10.9.0)
8. Explain the steps by which a Web is created and then used. (10.7.4)
9. How would you mine an email archive to identify and find people with interests similar to your own. (10.11.1)

Short-Essays and Hand-Worked Problems:

1. a) Research and describe in a paragraph how Arabic writing is presented in Unicode.
   b) Create page of Arabic characters from Unicode on your laser printer. (10.1.1)
2. Some people have bemoaned the fact that reading will be less important as interactive multimedia systems become more common. Handwriting letters is much less common now that people can communicate by telephone and email Do you think there is a connection between literacy and intelligence? Justify your answer. (10.2.2)
3. Why is linear presentation of text often easier to understand than a hypertext like presentation? (2.6.0, 10.2.4)
4. Get the cooperation of a friend and watch his/her eyes as they read a page of text. About how many times do their eyes stop? Describe a cognitive model of readability based on these observations. How is this related to the number of lines on the page? (10.2.4)
5. One simple measure of readability might be based the average length of words. Describe some other simple measures. How could you validate these measures? (10.3.1)
6. Calculate reading ease score for the sentence "The quick brown fox jumped over the lazy dog". (10.3.1)
7. Estimate how many words are written by the world's population in a day. Justify your estimate. (10.3.1)
8. Give an example of how people adapt their communication to the capabilities of the communication channels. (10.3.2)
9. Calculate the edit distance between "apple" and "able". (10.4.1)
10. Sketch parse trees for the following sentences (10.4.2, -A.5.4)
    a. The quick brown fox jumped over the lazy dog.
    b. Surely sweet Susan sells sea shells by the sea shore.
    c. Buffalo buffalo buffalo Buffalo buffalo. (Hint: Buffalo is a city, a verb, and an animal.)
11. Develop rules for text categorization of news stories dealing with the wheat crop harvest. (10.6.1)
12. Choose two Web search engines and compare their search results for several terms. How do you explain the differences? (10.7.4)
13. Are databases search engines? (3.9.0, 10.7.4)
14. How is Web searching similar to or different from a conversation with another person? (6.4.0, 10.7.4)
15. Describe some ways that document structure could support retrieval? ((sec:structure), 10.9.0)
16. When is a controlled vocabulary (such as a thesaurus) better than full text descriptions for retrieving information resources? (2.2.2, 10.9.0)
17. Distinguish between text retrieval and text data mining. (10.5.0, 10.9.0)
18. Explain the advantages and limitations of "proximity search". (10.9.1)
19. What is the ordering for a set of relevant documents? (10.9.2):

Total in collection	100	100	100
Relevant in collection	20	20	20
Number Retrieved	10	30	60
Relevant and Retrieved	4	7	9
Precision			
Recall			

20. Identify and describe the differences between two representations for text retrieval systems. (10.9.2)
21. Evaluate a Web search engine on several dimensions such as the interface and the accuracy. (10.9.5)
22. Explain the relationship between hubs, authorities, and the PageRank algorithm. (10.10.2)
23. When searching an index for a set of query terms that appear in a document set, why is it helpful to make the least frequent term in the set, the fist one to be searched? (10.10.2)
24. Classify the following questions according to the taxonomy in Fig. 3.8 (10.12.0)
    a.
    b.
25. In what way might the widespread use of English on the Web represent a network effect? (8.7.2, 10.13.0)
26. Give some examples of shallow methods for text processing (10.13.1)
27. Explain the distinction between shallow and deep linguistic processing methods for machine translation.(10.4.0,10.13.1)
28. Describe examples of algorithmic processing and statistical methods for text processing. Contrast the advantages and disadvantages these two approaches. (10.13.1)

Practicum:

1. Text processing.

Going Beyond:

1. Use a scanner to create a bmp file from a text document. Write programs to (a) distinguish lines of text from images and (b) Identify the title page. (10.1.5)
2. Develop a program to do table extraction from a bmp file. (10.1.6)
3. For people, what is the connection between ability in reading and ability in conversation. (6.4.0, 10.2.0)
4. Develop models for (a) reading and (b) writing that are consistent with limitations of cognitive processes we have discussed. ((sec:cognitiveprocessing), 10.2.0)
5. Build a filter for blocking articles having to do with automobiles from being displayed on a Web browser. (10.3.2)
6. Should we rely on Wikipedia for accurate information? (10.3.2)
7. Develop parse trees for the following sentences (10.4.2, -A.5.4)
   a.
   b.
   c.
8. Write a program to categorize news articles. (10.6.1)
9. Evaluate the usability of a legal search service. (7.10.2, 10.9.0)
10. Obtain a freeware search engine from the Web. Install it. Test it. (10.9.5)
11. If you were developing a Web robot to create an index on a specific topic, what would do to focus the search? (10.9.5)
12. If we ever develop effective question-answering systems; would people still use search engines? (10.9.5, 10.12.0)
13. Are there some types questions that a question-answering system should leave to a human being? Give some examples. Could this automatically determined? (10.12.0)
14. How are text similarity matching algorithms similar to copy detection algorithms? (8.2.5, 10.13.1)
15. Develop a system that determines the language in which a Web page is written. (10.13.1)
16. Given the variation of concepts across languages, is it ever possible to get an exact translation from one language to another? (10.13.1)

Teaching Notes

Objectives and Skills: The student will understand the basic components of a search engine as well as basic techniques for processing natural language.

Instructor Strategies: Some sections such as 10.6.1 are fairly specialized and may be dropped for some students.

Related Books

- Basbanes, N.A. *A Splendor of Letters: The Permanence of Books in an Impermanent World.* Harper-Collins, New York, 2004.
- Coulmus, F. *Writing Systems of the World.* Blackwell Press, Oxford UK, 1991.
- National Center for Reading *Report of the National Reading Panel: Teaching Children to Read.* National Institutes of Health, Bethesda MD, 2001
- Manning, C.D., Raghavan, P., and Schutze. H. *Introduction to Information Retrieval.* U. Cambridge Press, New York, 2008.
- Sunstein, C. *Infotopia: How many Minds Produce Knowledge.* Oxford University Press, New York, 2006
- Tancer, W. *Click: What Millions of People Are Doing Online and Why it Matters.* Hyperion, New York, 2008.