Applications of Computational Linguistics
Pius ten Hacken, English Seminar, WS 2001/02

Text Summarization

Noah Bubenhofer
25 January 2002

Contents

Theory of computational Summarization

Techniques

Demonstration

Definition

A summary text is a derivative of a source text condensed by selection and/or generalization on important content.

Input factors: source form, subject
Purpose factors: audience, function
Output factors: summary format, style

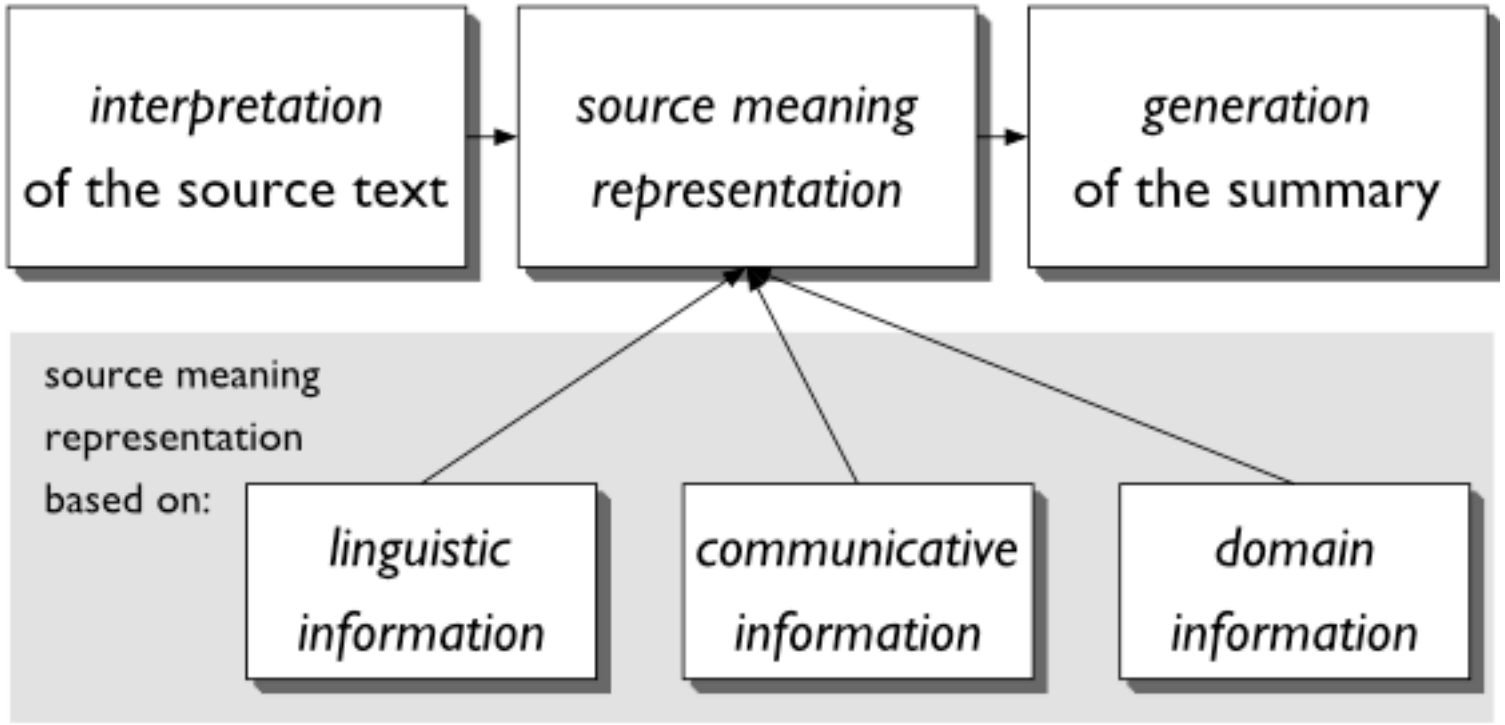
(Jones 7.4.1)

History

- 1950s: Luhn‘s auto-extracts statistical system
- 1970s: Domain based systems
- 1980s: Systems inspired by cognitive science theories
- 1990s: Use of different methods; domain-independent summarization

(Jones 7.4.0, Endres01: 456, Endres98: 297-365)

Process of Summarizing



Theory of computational Summarization

Techniques

Demonstration

Statistical techniques

- Length of sentence
- Indicators
  like „In conclusion...“, „We found...“
- Structure of paragraphs
  Beginnings and ends are important
- Key words
  Frequency of content words
- Acronyms

(Endres01: 460)

Domain Based Systems

- Input has always similar structure (e.g. scientific text)
- Statistical technique adapted to the domain
- Using „scripts“, „schemata“ or „templates“

(Endres98: 308)

Systems inspired by cognitive science

- Comparing input with information which is already in a „memory“
- Using „scripts“, „schemata“ or „templates“:
  description of a murder: murderer, victim, weapon

(Endres98: 310f)

Systems inspired by cognitive science

Example: SCISOR

- summarizes newspaper stories using a conceptual representation of knowledge about possible events

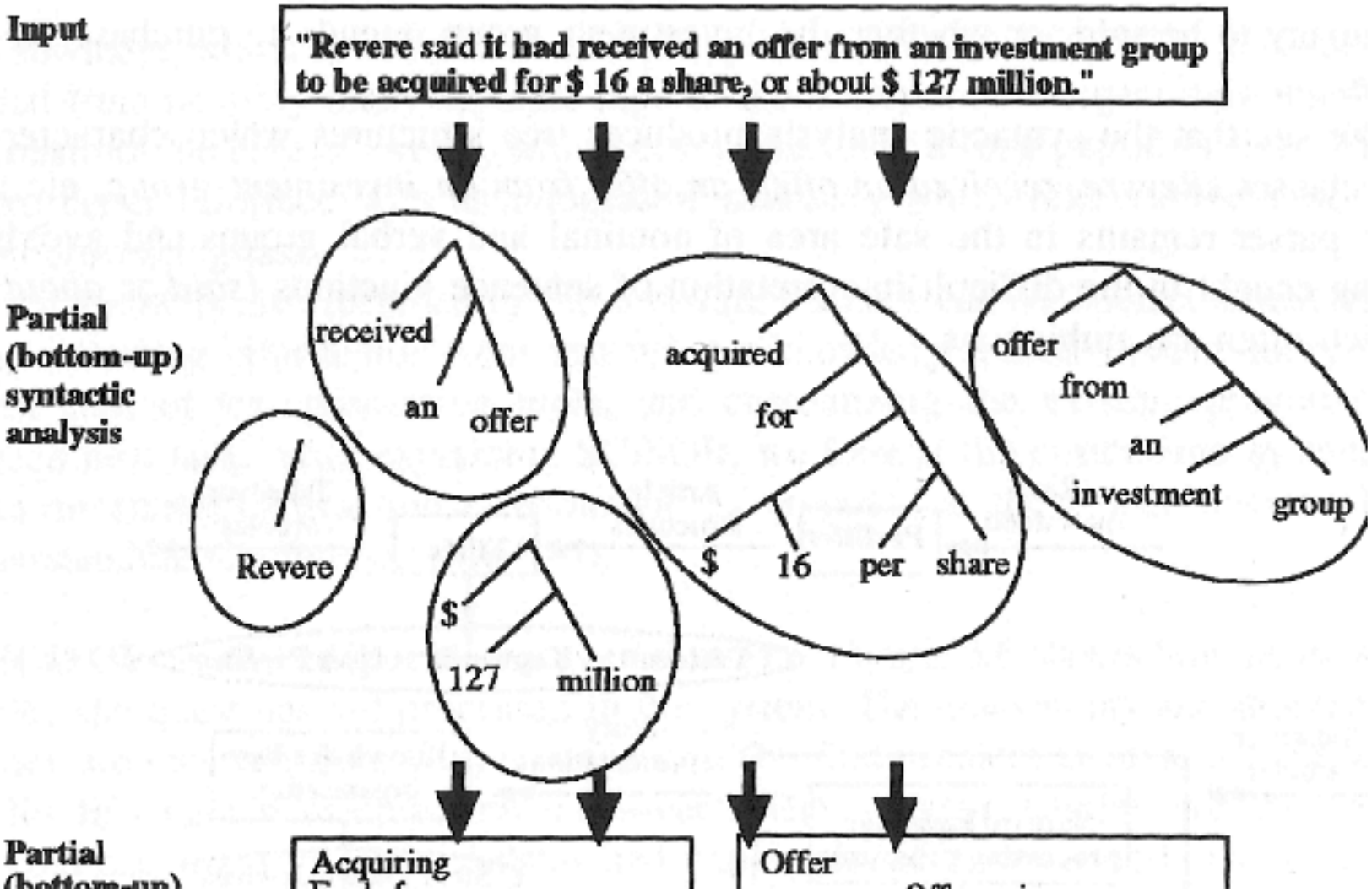
(Endres98: 319f)

Systems inspired by cognitive science Example: SCISOR

- memory contains schemata (e.g. about corporate merger)
- memory stores knowledge about whole episodes (e.g. all news items about a merger)

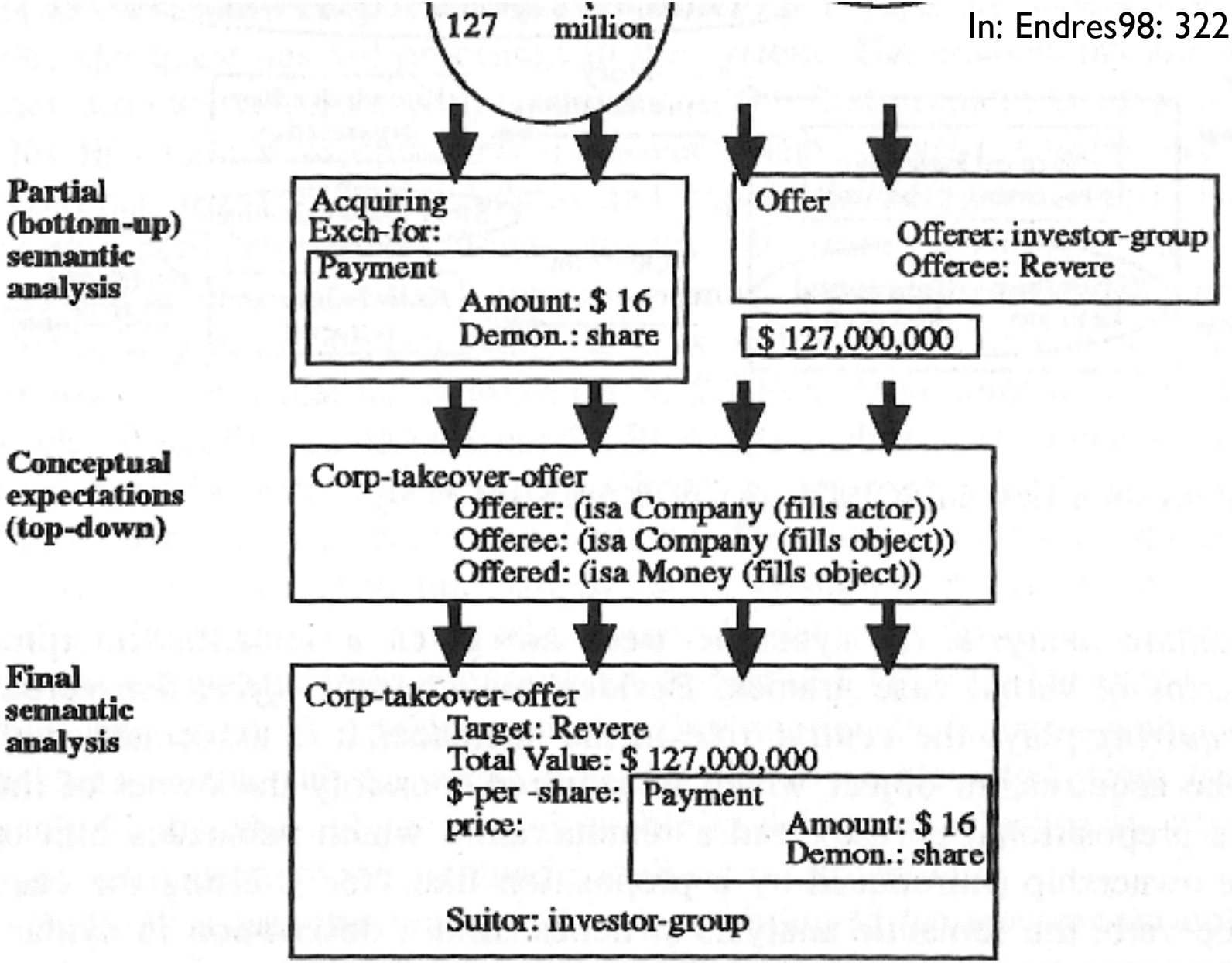
(Endres98: 319f)

Overview Text Summarization, Noah Bubenhofer, January 2002



In: Endres98: 322

In: Endres98: 322



Domain independent systems

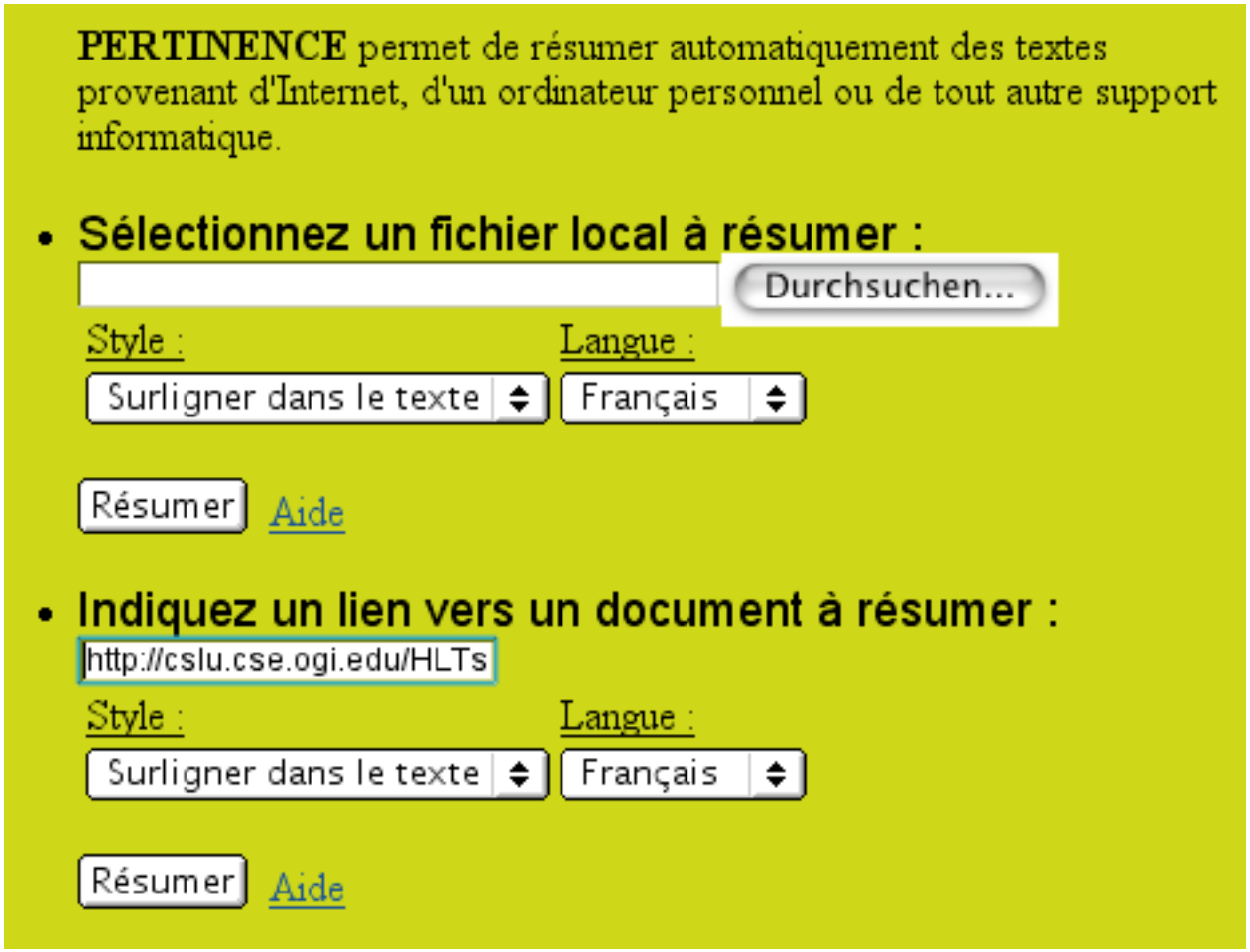
- mix of methods
- mostly statistical/linguistic approach

Domain independent systems

Example: Pertinence

Authors: Lehmam/Bouvet (F)
Input: text (txt, html, pdf, rtf, doc)
Output: summarization
http://www.pertinence.net (accessible via web)

Method: linguistic/statistical approach


PERTINENCE permet de résumer automatiquement des textes provenant d'Internet, d'un ordinateur personnel ou de tout autre support informatique.
Sélectionnez un fichier local à résumer :
Durchsuchen...
Style :
Langue :
Surligner dans le texte
Français
Résumer
Aide
Indiquez un lien vers un document à résumer :
http://cslu.cse.ogi.edu/HLTs
Style :
Langue :
Surligner dans le texte
Français
Résumer
Aide

Overview Text Summarization, Noah Bubenhofer, January 2002

formation of the *summary representation* using the source one and subsequent synthesis of the summary text. »Thus logical model emphasizes the role of text representations and the central transformation stage. »It thus focuses on what source representations should be like for summarizing, and on what condensation on important content requires. »Previous approaches to summarizing can be categorized and assessed, and new ones designed, according to (a) the nature of their source representation, including its distance from the source text, its relative emphasis on *linguistic*, *communicative* or *domain information* and therefore the structural model it employs and the way this marks important content; and (b) the nature of its processing steps, including whether all the model stages are present and how independent they are.»

7.4.2 Past Work

»

«For instance, reviewing past work (see [Pai90,SJ93]), source text extraction using statistical cues to select key sentences to form summaries is taking both source and summary texts as their own linguistic representations and also essentially conflating the interpretation and generation steps. »Approaches using cue words as a base for sentence selection are also directly exploiting only linguistic information for summarizing. »When headings or other locational criteria are exploited, this involves a very shallow source text representation depending on primarily linguistic notions of text grammar, though [L+93] has a richer grammar for a specific text type.»

«Approaches using scripts or frames on the other hand [YH85,DeJ79] involve deeper representations and ones of an explicitly domain-oriented kind motivated by properties of the world. »DeJong's work illustrates the case where the source representation is deliberately designed for summarizing, so there is little transformation effort in deriving the summary template representation. »In the approach of [Rau88], however, the hierarchic domain-based representation allows generalization for summarizing.»

«There has also been research combining different information types in representation. »Thus [Hah90] combines linguistic theme and domain structure in source representations, and seeks salient concepts in these for summaries.»

«Overall in this work, source reduction is mainly done by selection: this may use general, application-independent criteria, but is more commonly domain-guided as in [MHG84], or relies on prior, inflexible specification of the kind of information sought, as with [DeJ79], which may be as tightly constrained as in MUC. »There is no significant condensation of input content taken as a whole: in some cases even little length reduction. »There has been no systematic comparative study of different types of source representation for summarizing, or of context factor implications. »Work hitherto has been extremely fragmentary and, except where it resembles indexing or is for very specific and restricted kinds of material, has not been very successful. »The largest-scale automatic summarizing experiment done so far has been DeJong's, applying script-based techniques to news stories. »There do not appear to be any operational summarizing systems.»

7.4.3 Relevant Disciplines

»

Domain independent systems

Example: Extractor

Interactive Information Group, NRCC (CA)
Input: text (html, txt)
Output: list of key words, list of key sentences
http://extractor.iit.nrc.ca/

Method: linguistic/statistical approach

Extractor Demo

Extract Keyphrases from a Web Page

URL to Process:

http:// Extract Clear

- **Put the power of Extractor into your browser.**
- This demo of Extractor works with monolingual **English, French, Japanese, German, Spa** pages.
- Extractor will guess the language and the character encoding.
- Enter the URL of a web page for Extractor to process.
- Only the HTTP protocol is accepted.
- You can also copy and paste text from your web browser or word processor.

Overview Text Summarization, Noah Bubenhofer, January 2002

Keyphrases:

- summaries ○ Good ○ Bad
- representation ○ Good ○ Bad
- linguistics ○ Good ○ Bad
- interpretation ○ Good ○ Bad
- Extracting Information ○ Good ○ Bad
- exploiting ○ Good ○ Bad
- context ○ Good ○ Bad

To rate the quality of the keyphra
choose Good or Bad, then click he

Highlights:

- Next: 7.5 Computer Assistance in Text Creation and Editing Up: 7 Document Processing Pr
  Interpretation: **Extracting Information**
- The global process model has two major phases: **interpretation** of the source text involving
  analysis and integration of sentence analyses into an overall **source meaning representatio**
  summary by formation of the summary representation using the source one and subsequent sy
  text.
- For instance, reviewing past work (see [Pai90,SJ93]), source text extraction using statistical
  sentences to **form summaries** is taking both source and summary texts as their own linguisti
  essentially conflating the interpretation and generation steps.
- Approaches using cue words as a base for sentence selection are also **directly exploiting** o
  for summarizing.
  There has been no systematic comparative study of different types of source representation f

Theory of computational Summarization

Techniques

Demonstration

Test: Our Text

7.4 Summarization // Karen Sparck Jones // University of Cambridge, Cambridge, UK // Automatic abstracting was first attempted in the 1950s, in the form of Luhn's auto-extracts, (cf. [Pai90]); but since then there has been little work on, or progress made with, this manifestly very challenging task. However the increasing volume of machine-readable text, and advances in natural language processing, have stimulated a new interest in automatic summarizing reflected in the 1993 Dagstuhl Seminar, Summarizing text for intelligent communication [ENHSJ95]. Summarizing techniques tested so far have been limited either to general, but shallow and weak approaches, or to deep but highly application-specific ones. There is a clear need for more powerful, i.e., general but adaptable, methods. But these must as far as possible be linguistic methods, not requiring extensive world knowledge, and ones able to deal with large-scale text structure as well as individual sentences. [...]

Test: Result (Extractor)

Keywords

- form summaries
- source meaning representation
- linguistics supplying
- directly exploiting
- context
- discourse
- analytical framework covering

Test: Result (Pertinence/Extractor)

Pertinence/Both/Extractor (1/4)

- Work done hitherto, relevant technologies, and required directions for new research are usefully characterized by reference to an analytical framework covering both factors affecting summarizing and the essential summarizing process.
- I shall concentrate on text, but the framework applies to **discourse** in general including dialogue.
- The global process model has two major phases: interpretation of the source text involving both local sentence analysis and integration of sentence analyses into an overall source meaning representation; and generation of the summary by formation of the summary representation using the source one and subsequent synthesis of the summary text.

Pertinence/Both/Extractor (2/4)

- Previous approaches to summarizing can be categorized and assessed, and new ones designed, according to (a) the nature of their source representation, including its distance from the source text, its relative emphasis on linguistic, communicative or domain information and therefore the structural model it employs and the way this marks important content; and (b) the nature of its processing steps, including whether all the model stages are present and how independent they are.
- For instance, reviewing past work (see [Pai90,SJ93]), source text extraction using statistical cues to select key sentences to form summaries is taking both source and summary texts as their own linguistic representations and also essentially conflating the interpretation and generation steps.

Pertinence/Both/Extractor (3/4)

- Approaches using cue words as a base for sentence selection are also **directly exploiting** only linguistic information for summarizing.
- There has been no systematic comparative study of different types of source representation for summarizing, or of **context** factor implications.

Pertinence/Both/Extractor (4/4)

- These include human summarizing, a trained professional skill that provides an iterative, processual view of summarizing often systematically exploiting surface cues; discourse and text linguistics supplying a range of theories of discourse structure and of text types bearing on summarizing in general, on different treatments suited to different source types, and on the relation between texts, as between source and summary texts; work on discourse comprehension, especially that involving or facilitating summarizing; library and information science studies of user activities exploiting abstracts e.g., to serve different kinds of information need; research on user modeling in text generation, for tailoring summaries; and NLP technology generally in supplying both workhorse sentence processing for interpretation and generation and methods for dealing with local coherence, as well as results from experiments with forms of large-scale text structure, if only for generation so far, not recognition.

Test 2: Our Text (German)

Verständigungsschwierigkeiten im globalen Dorf

Englisch als «lingua franca» in Wirtschaftsbetrieben

Von Ulla Kleinberger Günther*

Dank der wirtschaftlichen Globalisierung ist der Siegeszug des Englischen auch in Betrieben nicht mehr aufzuhalten. Entsprechend wird das Erlernen dieser Sprache verlangt und gefördert. Die Dominanz des Englischen im wirtschaftlichen Bereich entspricht jedoch oft nicht der wirklichen Sprachsituation. Dort gilt es, verschiedenste Formen der Kommunikation zu unterscheiden, und nicht für jede ist das Englische geeignet. Die Mehrsprachigkeit der Schweiz ist ein nicht zu unterschätzender Vorteil.

Englisch wird in vielen beruflichen Bereichen der Industrieländer als die wichtigste «lingua franca» der Gegenwart angesehen. Als Zweitsprache dient sie sowohl mündlich wie schriftlich der interkulturellen Kommunikation. In der Schule und der Aus- und Weiterbildung in der Deutschschweiz werden die Leute...

NZZ, 12. 1. 02, p 83

Test 2: Results (Keywords)

Extractor

Englisch
wirtschaftlichen Globalisierung
Kommunikation
interkulturellen Kommunikation
beruflichen Alltag
betrieblichen Alltag
Mitarbeitern

Test 2: Results (Keywords)

Extractor

Englisch
wirtschaftlichen Globalisierung
Kommunikation
interkulturellen Kommunikation
beruflichen Alltag
betrieblichen Alltag
Mitarbeitern

Statistical Approach

min 3x

Alltag
Anforderungen
Bereich
Englisch
Forderung
Französisch
Italienisch
Kommunikation
Kompetenzen
Mehrsprachigkeit
Mitarbeitern
Rolle
Situation
Spanisch
Sprachen
Sprachkompetenz
Umfeld
Umgang

Test 2: Results (Keywords)

Extractor

Englisch
wirtschaftlichen Globalisierung
Kommunikation
interkulturellen Kommunikation
beruflichen Alltag
betrieblichen Alltag
Mitarbeitern

Statistical Approach

Alltag
Anforderungen
Bereich
Englisch
Forderung
Französisch
Italienisch
Kommunikation
Kompetenzen
Mehrsprachigkeit
Mitarbeitern
Rolle
Situation
Spanisch
Sprachen
Sprachkompetenz
Umfeld
Umgang

min 3x
min 5x

Test 2: Results (Keywords)

Extractor

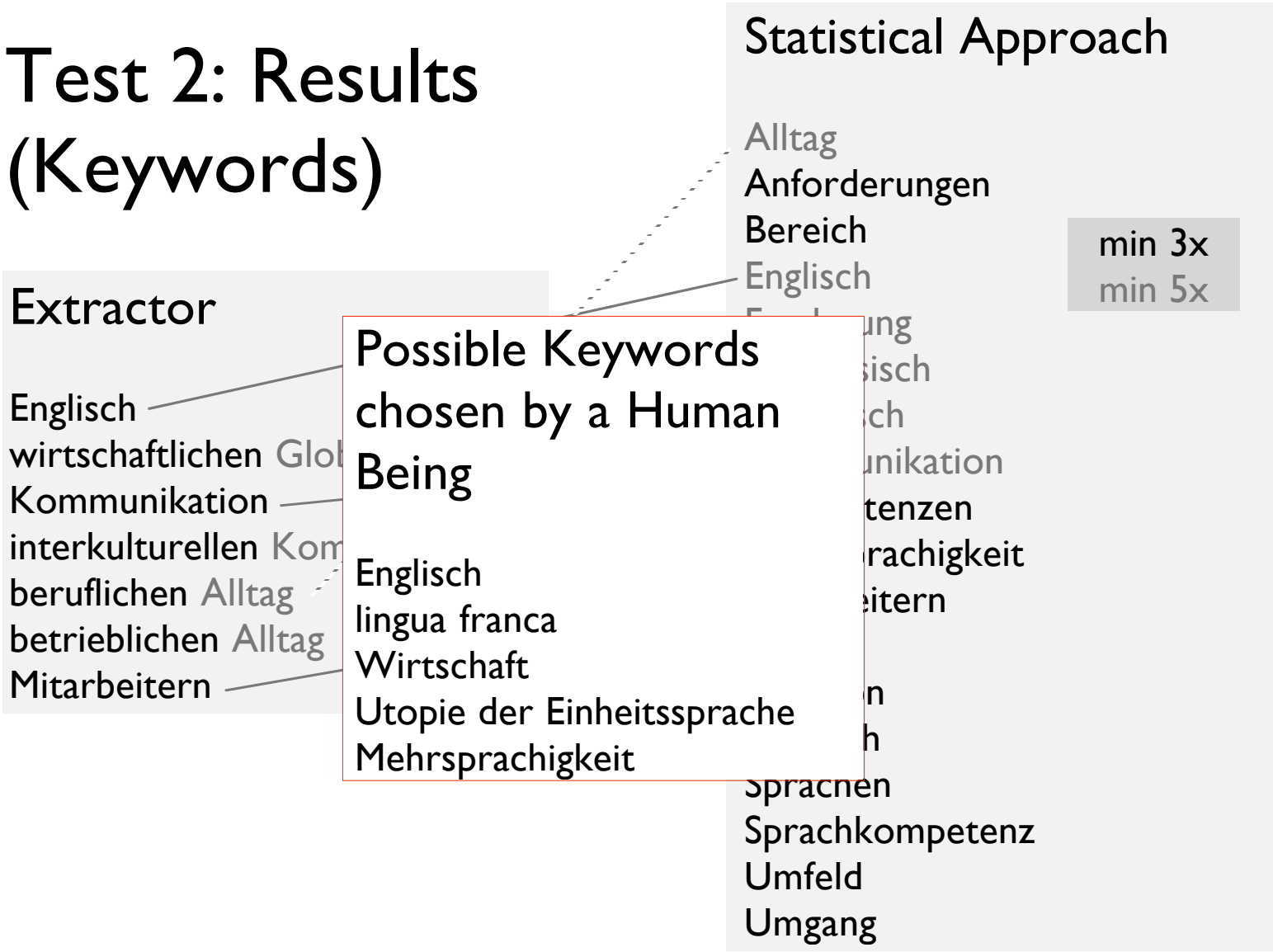
Englisch
wirtschaftlichen Globalisierung
Kommunikation
interkulturellen Kommunikation
beruflichen Alltag
betrieblichen Alltag
Mitarbeitern

Statistical Approach

Alltag
Anforderungen
Bereich
Englisch
Forderung
Französisch
Italienisch
Kommunikation
Kompetenzen
Mehrsprachigkeit
Mitarbeitern
Rolle
Situation
Spanisch
Sprachen
Sprachkompetenz
Umfeld
Umgang

min 3x
min 5x

Test 2: Results (Keywords)



Test 2: Results Extractor („Highlights“) 1/2

- Dank der **wirtschaftlichen Globalisierung** ist der Siegeszug des Englischen auch in Betrieben nicht mehr aufzuhalten.
- Dort gilt es, verschiedenste Formen der **Kommunikation** zu unterscheiden, und nicht für jede ist das Englische geeignet.
- **Englisch** wird in vielen beruflichen Bereichen der Industrieländer als die wichtigste «lingua franca» der Gegenwart angesehen.
- Als Zweitsprache dient sie sowohl mündlich wie schriftlich der **interkulturellen Kommunikation.**
- Die Forderung der Wirtschaft nach einer einheitlichen Sprachform ist hinsichtlich des Effizienzgedankens - temporal, monetär usw. - durchaus verständlich, im **beruflichen Alltag** stellt sich die Situation jedoch keineswegs so klar und eindeutig dar.
- [...]

Test 2: Results Extractor („Highlights“) 2/2

- Auf Grund von Ergebnissen aus Befragungen von **Mitarbeitern** in sechs international tätigen Schweizer Grossbetrieben muss die Rolle des Englischen im beruflichen Alltag differenzierter betrachtet werden; die Auseinandersetzung darf also keineswegs bei der pauschalen Forderung nach Englisch in allen Bereichen und für alle Begebenheiten belassen werden, sondern sie muss auf Grund empirisch erhobener Daten weitergeführt und entsprechend differenziert betrachtet werden.
- Diese Erkenntnisse aus dem **betrieblichen Alltag** haben für die Sprachpolitik Bedeutung

Conclusion

- There is a need for automatic summarization (information retrieval)
- Automatic summarization is successful especially where it is limited to defined domains
- Much more improvement is possible:
  better linguistic analysis (syntax and semantics)
  better cognitive science based methods
- Fundamental techniques of computer linguistics are needed:
  syntax parsing, semantical analysis etc.

Appendix

References

- Endres-Niggemeyer, Brigitte (1998): Summarizing Information. Springer, Berlin, Heidelberg, New York.
- Endres-Niggemeyer, Brigitte (2001): Textzusammenfassung. In: Carstensen, K.-U. (ed.): Computerlinguistik und Sprachtechnologie. Eine Einführung. Spektrum, Heidelberg, Berlin.
- Endres-Niggemeyer, Brigitte: Website: www.ik.fh-hannover.de/ik/personen/ben/ben.htm
- Mani, Inderjeet (2001): Automatic Summarization. John Benjamins, xi+285pp, paperback ISBN 1-58811-060-5, Natural Language Processing, 3.
- Spack Jones, Karen (1997): Summarization. In: Cole/Mariani (eds.): Survey of the State of the Art in Human Language Technology. Cambrige Univ. Press. (http://cslu.cse.ogi.edu/hltsurvey)

Web: www.summarization.com (by Dragomir Radev)

Summarization Systems I

- **Pertinence:**
  Lehmam/Bouvet (F)
  Output: summarization
  http://www.pertinence.net
- **Extractor:**
  Interactive Information Group, NRCC (CA)
  Output: list of key words, list of key sentences
  http://extractor.iit.nrc.ca/

Summarization Systems 2

- MEAD:
  Center for Language and Speech Processing, Hopkins University, Baltimore (Dragomir Radev)
  Automatic Summarization of Multilingual Documents
  http://www.clsp.jhu.edu/ws2001/groups/asmd/
- Inxight Summarizer:
  Inxight Software Inc., Santa Clara, California
  Output: Abstract
  http://www.inxight.com/

Summarization Systems 3

- Scisor:
  Jacobs/Rau, 1990
  Automatic summarization of news about mergers in economy

- STREAK:
  McKeown et al., 1995
  Generates short descriptions of basketball games out of a database

Endres-Niggemeyer: 457-458

Summarization Systems 4

- Configurable Text Summarization System:
  Barker/Chali/Copeck/Matwin/Szpakowicz 1998
  University of Ottawa
  Input: Query on texts
  Output: Summarized text

CTSS Process