Source Encoding and Compression

Jukka Teuhola
University of Turku
Dept. of Information Technology

Spring 2016

General

- **Self-study course, starting lecture:** 02.03.2016
- **Extent:** 5 sp (3 cu)
- **Level:** Advanced
- **Preliminary knowledge:** Data structures and algorithms I, basics of probability calculus
- **Material:** Lecture notes and Powerpoint slides available via the course homepage. No textbook is needed.
- **Homework:** 10 small exercise tasks will be given. Solutions must be submitted to the lecturer before taking the examination. Minimum: 5 solutions acceptably solved.
- **Examinations:** Three attempts; May 10, Jun 13, Sep ?

Optional literature

- T. C. Bell, J. G. Cleary, I. H. Witten: *Text Compression*, 1990.
- R. W. Hamming: *Coding and Information Theory*, 2nd ed., Prentice-Hall, 1986.
- K. Sayood: *Introduction to Data Compression*, 3rd ed., Morgan Kaufmann, 2006.
- K. Sayood: *Lossless Compression Handbook*, Academic Press, 2003.
- I. H. Witten, A. Moffat, T. C. Bell: *Managing Gigabytes*: compressing and indexing documents and images, Morgan Kaufmann, 1999.
- Miscellaneous articles

Contents

1. Basic concepts

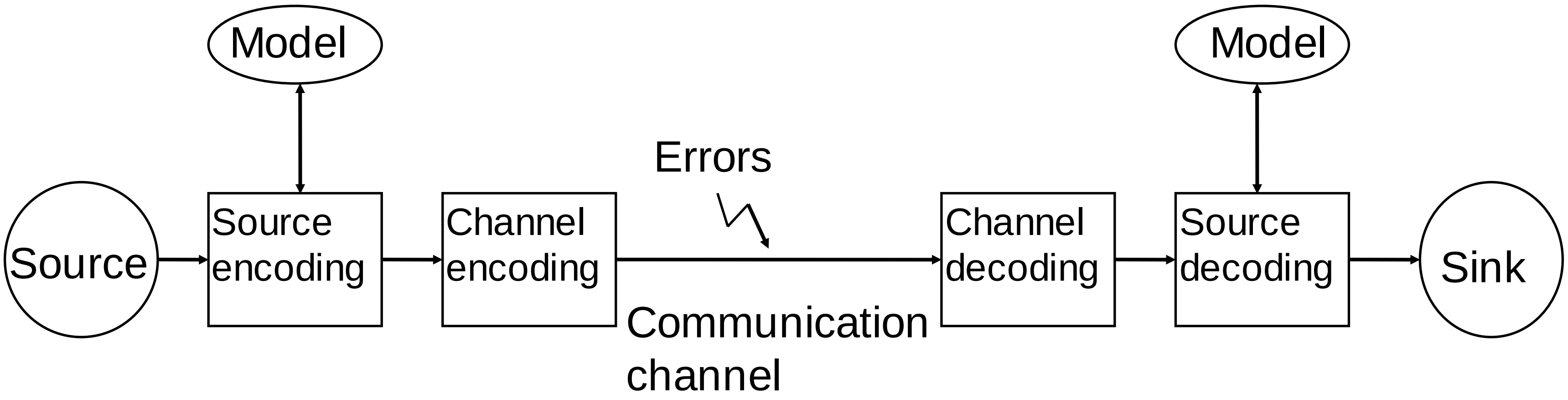
- **Data compression:**
  - Minimize the size of information representation.
  - Reduce the *redundancy* of the original representation.

- **Purposes:**
  - Save storage space.
  - Reduce transmission time.

- **Basic approaches:**
  - *Lossless* compression: decompression into exactly the original form (typical for text).
  - *Lossy* compression: decompression into approximately the original form (typical for signals and images).

Basic concepts (cont.)

- **Fields of coding theory:**
  - *Source coding*: purpose to minimize the size
  - *Channel coding*: detection and correction of transmission errors.



  - Also: *cryptography*: Encryption of private/secret information

Basic concepts (cont.)

- **Phases of data compression:**
  - *Modelling* of the source
  - *Source encoding* (called also *entropy coding*), using the model
- **Other viewpoints:**
  - Speed of compression / decompression
  - Size of the model
- **Classification by lengths of coding units:**
  - *Fixed-to-fixed* coding
  - *Variable-to-fixed coding*
  - *Fixed-to-variable coding*
  - *Variable-to-variable coding*

Examples of models

1. Character distribution

Char	Prob
A	0.10
B	0.05
C	0.08
D	0.06
E	0.15
…..	…..

2. Successor distribution

Char	Succ	Prob
A	A	0.01
A	B	0.20
A	C	0.10
A	D	0.25
…..	…..	……
B	A	0.15
B	B	0.02
B	C	0.01
B	D	0.01
…..	…..	…..

3. Dictionary

Word	Prob
ALL	0.02
ALWAYS	0.01
ARE	0.05
AS	0.03
AT	0.02
BASIC	0.01
BEGIN	0.01
…..	…..

Basic concepts (cont.)

- **Main classes of text compression methods:**
  - *Dictionary* methods
  - *Statistical* methods
- **Classification based on availability of the source:**
  - *Off-line methods*
  - *On-line methods*
- **Classification based on the status of the model:**
  - *Static methods*
  - *Semiadaptive methods*
  - *Adaptive methods*
- **Measurement of compression efficiency:**
  - Compression ratio: Source size / compressed size
  - Bits per source symbol (character, pixel, etc.)

Illustration of a static method

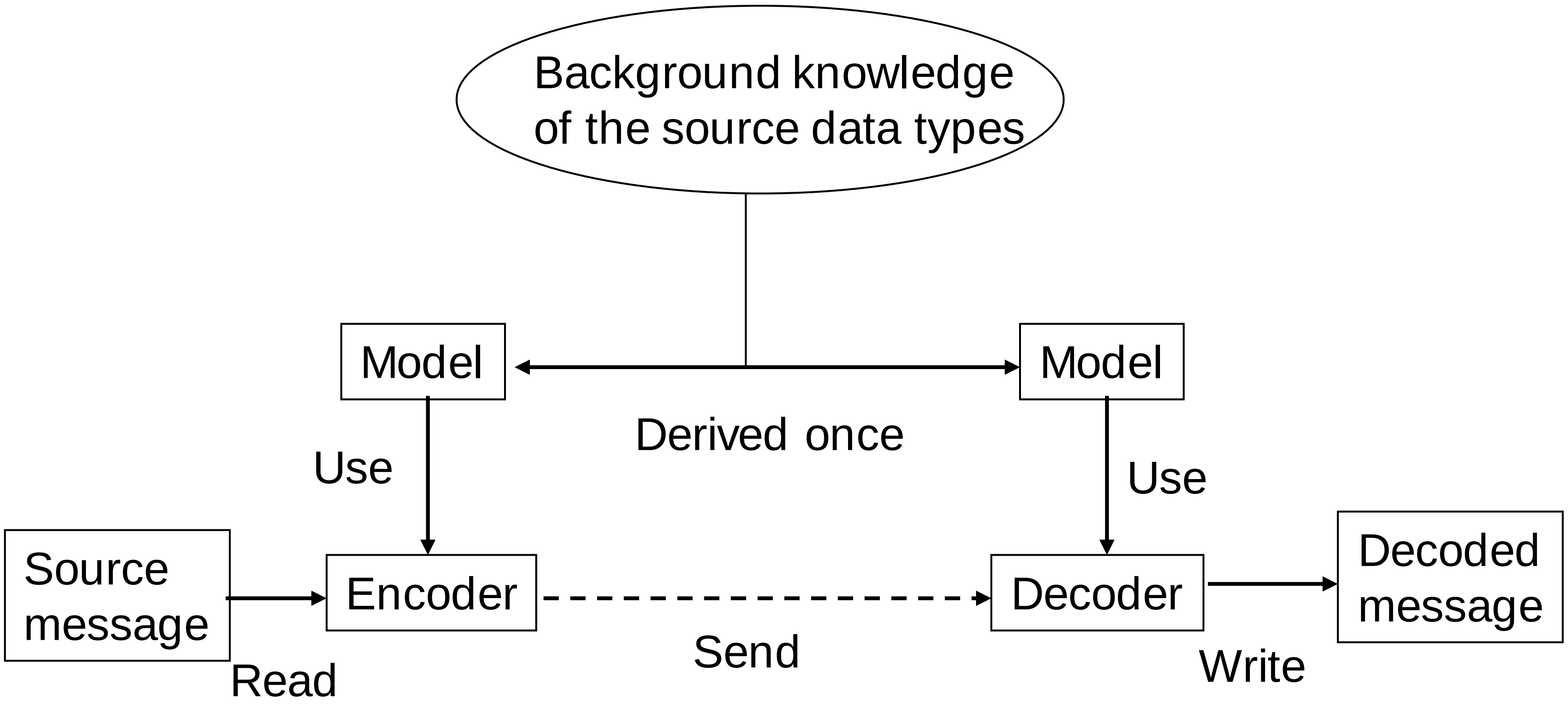


Illustration of a semiadaptive method

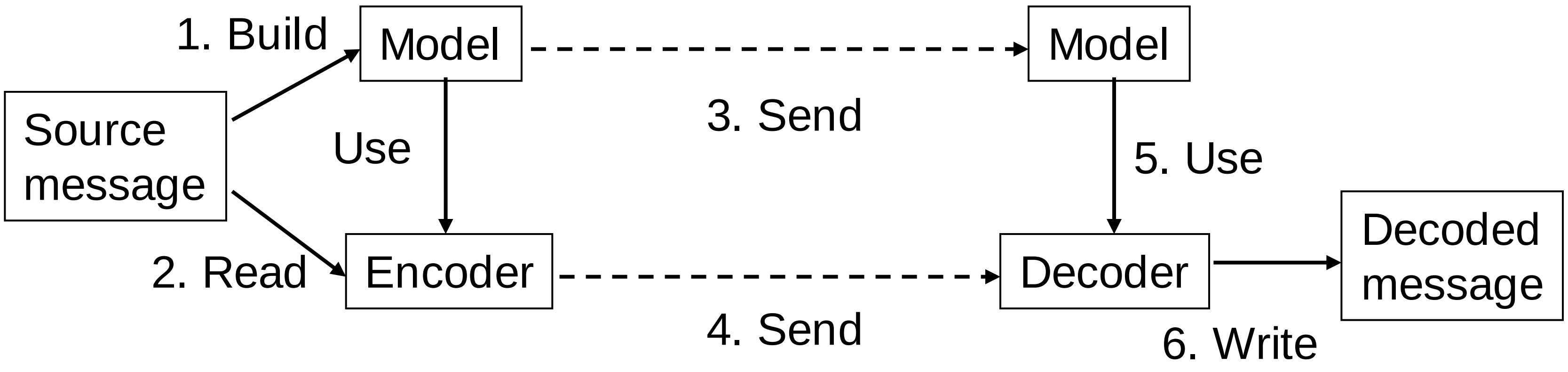
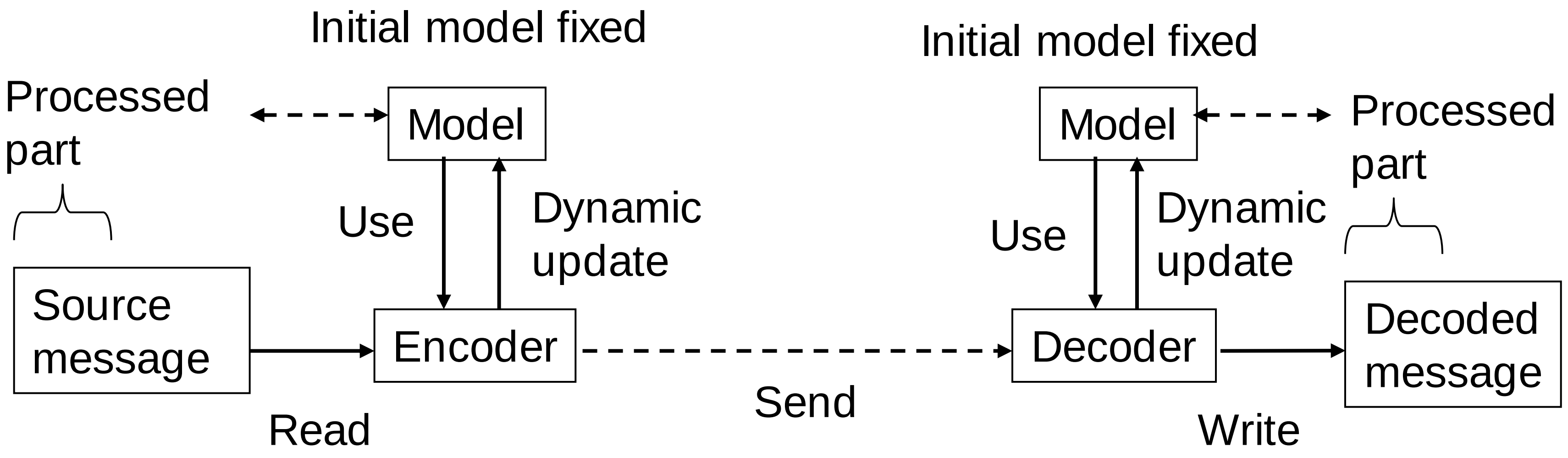


Illustration of an adaptive method

Models are updated dynamically, based on the already processed part of the source, known to both encoder and decoder.



Initial model fixed

Processed part

Model

Use

Dynamic update

Source message

Read

Encoder

Send

Decoder

Write

Initial model fixed

Model

Use

Dynamic update

Processed part

Decoded message