Statistical Machine Translation

What works and what does not

Andreas Maletti

Universität Stuttgart

maletti@ims.uni-stuttgart.de

Stuttgart — May 14, 2013

Main notions

Machine translation (MT)

Automatic natural language *translation* (by a computer)

as opposed to:

- manual translation
- computer-aided translation (e.g., translation memory)

Statistical machine translation (SMT)

MT using systems *automatically* obtained from (many) *translations*

as opposed to:

- rule-based machine translation (old) SYSTRAN
- example-based machine translation translation by analogy

Main notions

Machine translation (MT)

Automatic natural language *translation* (by a computer)

as opposed to:

- manual translation
- computer-aided translation (e.g., translation memory)

Statistical machine translation (SMT)

MT using systems *automatically* obtained from (many) *translations*

as opposed to:

- rule-based machine translation (old) SYSTRAN
- example-based machine translation translation by analogy

Short history

Timeline

1. Dark age (60s–90s)
   - rule-based systems (e.g., SYSTRAN)
   - CHOMSKYAN approach
   - perfect translation, poor coverage
2. Reformation (1991–present)
   - phrase-based and syntax-based systems
   - statistical approach
   - cheap, automatically trained
3. Potential future
   - semantics-based systems (e.g., FRAMENET-based)
   - semi-supervised, statistical approach
   - basic understanding of translated text

Short history

Timeline

1. Dark age (60s–90s)
   - rule-based systems (e.g., SysTran)
   - Chomskyan approach
   - perfect translation, poor coverage
2. Reformation (1991–present)
   - phrase-based and syntax-based systems
   - statistical approach
   - cheap, automatically trained
3. Potential future
   - semantics-based systems (e.g., FrameNet-based)
   - semi-supervised, statistical approach
   - basic understanding of translated text

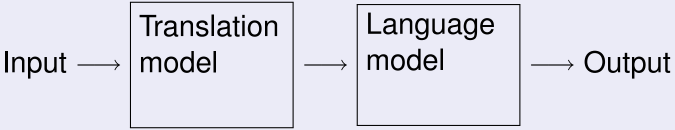
Short history

Timeline

1. Dark age (60s–90s)
   - rule-based systems (e.g., SYSTRAN)
   - CHOMSKYAN approach
   - perfect translation, poor coverage
2. Reformation (1991–present)
   - phrase-based and syntax-based systems
   - statistical approach
   - cheap, automatically trained
3. Potential future
   - semantics-based systems (e.g., FRAMENET-based)
   - semi-supervised, statistical approach
   - basic understanding of translated text

Standard pipeline

Schema



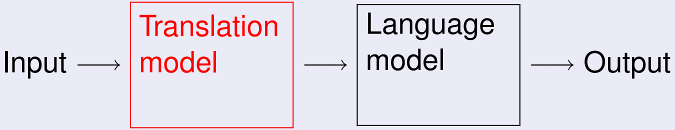
(the models are often integrated in practice)

Required resources

bilingual text (sentences in both languages)	1.5M sent.
monolingual text (in target language)	44M sent.

Standard pipeline

Schema



Input → Translation model → Language model → Output

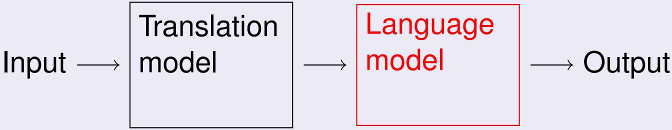
(the models are often integrated in practice)

Required resources

bilingual text (sentences in both languages)	1.5M sent.
monolingual text (in target language)	44M sent.

Standard pipeline

Schema



Input ⟶ Translation model ⟶ Language model ⟶ Output

(the models are often integrated in practice)

Required resources

- bilingual text (sentences in both languages) 1.5M sent.
- monolingual text (in target language) 44M sent.

Standard pipeline

Example (Source: GOOGLE translate)

- Input:

  What works and what does not

- Segmentation:

  What works | *and what does not*

- Translation model output:

Was funktioniert	*und was nicht*
Was am	*und was nicht funktioniert*
Was funktioniert am	*und welche nicht*
	ist und was nicht

Standard pipeline

Example (Source: GOOGLE translate)

- Input:

  What works and what does not

- Segmentation:

What works	*and what does not*

- Translation model output:

Was funktioniert	*und was nicht*
Was am	*und was nicht funktioniert*
Was funktioniert am	*und welche nicht*
	ist und was nicht

Standard pipeline

Example (Source: GOOGLE translate)

- Input:

  What works and what does not

- Segmentation:

  | *What works* | *and what does not* |

- Translation model output:

Was funktioniert	*und was nicht*
Was am	*und was nicht funktioniert*
Was funktioniert am	*und welche nicht*
	ist und was nicht

Standard pipeline

Example (Source: GOOGLE translate)

- Input:

  What works and what does not

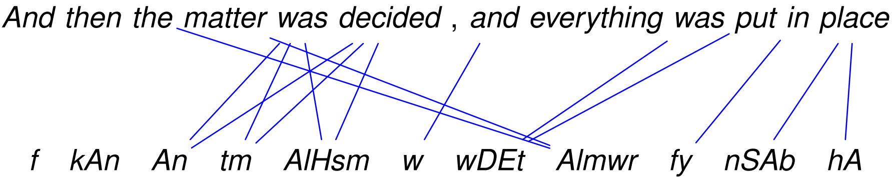
- Segmentation:

  | *What works* | *and what does not* |

- Translation model output:

Was funktioniert	*und was nicht*
Was am	*und was nicht funktioniert*
Was funktioniert am	*und welche nicht*
	ist und was nicht

Phrase-based machine translation


And then the matter was decided , and everything was put in place
f kAn An tm AlHsm w wDEt Almwr fy nSAb hA

Extracted information

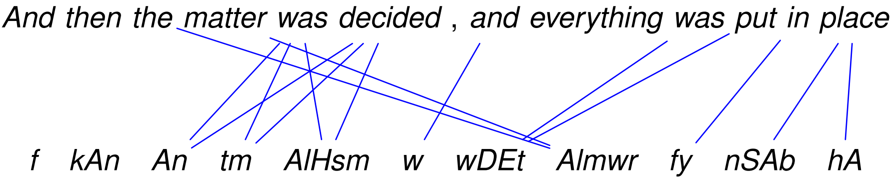
Segmentation:

And then the matter was decided , and everything was put in place

Phrase translation:

Reordering:

Phrase-based machine translation


And then the matter was decided , and everything was put in place
f kAn An tm AlHsm w wDEt Almwr fy nSAb hA

Extracted information

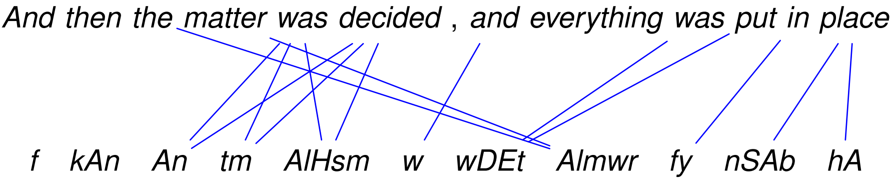
Segmentation:

And then the matter was decided , and everything was put in place

Phrase translation:

Reordering:

Phrase-based machine translation


And then the matter was decided , and everything was put in place
f kAn An tm AlHsm w wDEt Almwr fy nSAb hA

Extracted information

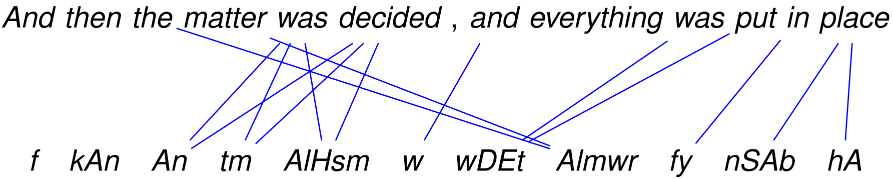
Segmentation:

And then $_1$ *the matter* $_2$ *was decided* $_3$ *, and everything* $_4$ *was put* $_5$ *in place* $_6$

Phrase translation:

Reordering:

Phrase-based machine translation



And then the matter was decided , and everything was put in place

f kAn An tm AlHsm w wDEt Almwr fy nSAb hA

Extracted information

Segmentation:

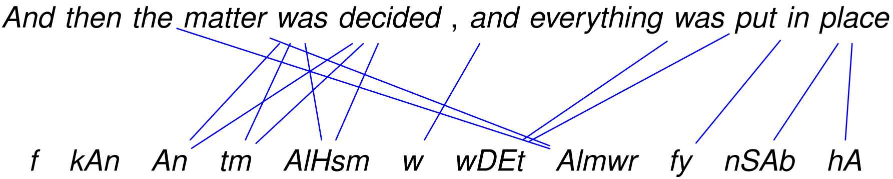
$\boxed{\textit{And then}}_1$ $\boxed{\textit{the matter}}_2$ $\boxed{\textit{was decided}}_3$ $\boxed{\textit{, and everything}}_4$ $\boxed{\textit{was put}}_5$ $\boxed{\textit{in place}}_6$

Phrase translation:

$\boxed{\textit{f kAn}}_1$ $\boxed{\textit{Almwr}}_2$ $\boxed{\textit{An tm AlHsm}}_3$ $\boxed{\textit{w}}_4$ $\boxed{\textit{wDEt}}_5$ $\boxed{\textit{fy nSAb hA}}_6$

Reordering:

Phrase-based machine translation



And then the matter was decided , and everything was put in place
f kAn An tm AlHsm w wDEt Almwr fy nSAb hA

Extracted information

Segmentation:

$\boxed{\textit{And then}}_1$ $\boxed{\textit{the matter}}_2$ $\boxed{\textit{was decided}}_3$ $\boxed{\textit{, and everything}}_4$ $\boxed{\textit{was put}}_5$ $\boxed{\textit{in place}}_6$

Phrase translation:

$\boxed{\textit{f kAn}}_1$ $\boxed{\textit{Almwr}}_2$ $\boxed{\textit{An tm AlHsm}}_3$ $\boxed{\textit{w}}_4$ $\boxed{\textit{wDEt}}_5$ $\boxed{\textit{fy nSAb hA}}_6$

Reordering: (1 3 4 5 2 6)

How it works

Technical talks

Marion Weller	phrase-based MT
Daniel Quernheim and Nina Seemann	syntax-based MT

Small players

Research at IMS

- Phrase-based MT (head: Dr. Alexander Fraser)
  - Fabienne Braune
  - Fabienne Cap
  - Anita Ramm
  - Marion Weller
- Syntax-based MT (head: Dr. Andreas Maletti)
  - Fabienne Braune
  - Daniel Quernheim
  - Nina Seemann

Small players

Research at IMS

- Phrase-based MT (head: Dr. Alexander Fraser)
  - Fabienne Braune
  - Fabienne Cap
  - Anita Ramm
  - Marion Weller
- Syntax-based MT (head: Dr. Andreas Maletti)
  - Fabienne Braune
  - Daniel Quernheim
  - Nina Seemann

Small players

Research at IMS

- Phrase-based MT (head: Dr. Alexander Fraser)
    - Fabienne Braune
    - Fabienne Cap
    - Anita Ramm
    - Marion Weller
- Syntax-based MT (head: Dr. Andreas Maletti)
    - Fabienne Braune
    - Daniel Quernheim
    - Nina Seemann

Big players

Commercial systems

Asia Online	Language Studio
Google	Google translate
IBM	WebSphere Translation Server
Microsoft	Bing translator
SAIC	Omnifluent
SDL Because Business is Global	
SYSTRAN	
…	

Big players

Commercial systems

Asia Online	Language Studio
Google	GOOGLE translate
IBM	WebSphere Translation Server
Microsoft	BING translator
SAIC	OMNIFLUENT
SDL Because Business is Global	
SYSTRAN	
…	

Soon also ebay

Failures

Failures

Applications

- Technical manuals

Example (An mp3 player)

The synchronous manifestation of lyrics is a procedure for can broadcasting the music, waiting the mp3 file at the same time showing the lyrics.

With the this kind method that the equipments that synchronous function of support up broadcast to make use of document create setup, you can pass the LCD window way the check at the document contents that broadcast.

That procedure returns offerings to have to modify, and delete, and stick top , keep etc. edit function.

Failures

Applications

- Technical manuals

Example (An mp3 player)

The synchronous manifestation of lyrics is a procedure for can broadcasting the music, waiting the mp3 file at the same time showing the lyrics.

With the this kind method that the equipments that synchronous function of support up broadcast to make use of document create setup, you can pass the LCD window way the check at the document contents that broadcast.

That procedure returns offerings to have to modify, and delete, and stick top , keep etc. edit function.

Failures

Applications

- Technical manuals

Example (An mp3 player)

The synchronous manifestation of lyrics is a procedure for can broadcasting the music, waiting the mp3 file at the same time showing the lyrics.

With the this kind method that the equipments that synchronous function of support up broadcast to make use of document create setup, you can pass the LCD window way the check at the document contents that broadcast.

That procedure returns offerings to have to modify, and delete, and stick top , keep etc. edit function.

Failures

Applications

- Technical manuals
- tripadvisor

Example (Hotel Uppsala, Sweden)

Wir hatten die Zimmer eingestuft wird als “Superior” weil sie renoviert wurde im letzten Jahr oder zwei. Unsere Zimmer hatten Parkettboden und waren sehr geräumig. Man musste allerdings nicht musste seitwärts bewegen.

Failures

Applications

- Technical manuals
- tripadvisor

Example (Hotel Uppsala, Sweden)

Wir hatten die Zimmer eingestuft wird als “Superior” weil sie renoviert wurde im letzten Jahr oder zwei. Unsere Zimmer hatten Parkettboden und waren sehr geräumig. Man musste allerdings nicht musste seitwärts bewegen.

— *We stayed in rooms classified as “superior” because they had been renovated in the last year or two. Our rooms had wood floors and were roomy. You didn’t have to walk sideways to move around.*

Failures

Applications

- Technical manuals
- tripadvisor
- US military

Example (Jones, Shen, Herzog 2009)

Soldier: Okay, what is your name?
Local: Abdul.
Soldier: And your last name?
Local: Al Farran.

Failures

Applications

- Technical manuals
- tripadvisor
- US military

Example (Jones, Shen, Herzog 2009)

Soldier: Okay, what is your name?
Local: Abdul.
Soldier: And your last name?
Local: Al Farran.

Speech-to-text machine translation

Soldier: Okay, what's your name?
Local: milk a mechanic and I am here I mean yes

Failures

Applications

- Technical manuals
- tripadvisor
- US military

Example (Jones, Shen, Herzog 2009)

Soldier: Okay, what is your name?
Local: Abdul.
Soldier: And your last name?
Local: Al Farran.

Speech-to-text machine translation

Soldier: Okay, what's your name?
Local: milk a mechanic and I am here I mean yes
Soldier: What is your last name?
Local: every two weeks my son's name is ismail

Failures

Applications

- Technical manuals
- tripadvisor
- US military
- MSDN, Knowledge Base
- …

But in many cases it actually works . . .

Selected application

Lecture translation

- KIT Karlsruhe Institute of Technology
- real-time speech-to-text machine translation
- combines automatic speech recognition and SMT
- requires lecturer training and terminology training
- automatically provides subtitles to lecture video

Video

`http://www.youtube.com/watch?v=x51L0wpr-88`

Selected application

Lecture translation

- KIT Karlsruhe Institute of Technology
- real-time speech-to-text machine translation
- combines automatic speech recognition and SMT
- requires lecturer training and terminology training
- automatically provides subtitles to lecture video

Video

`http://www.youtube.com/watch?v=x5lL0wpr-88`

Summary

SMT works well

- between similar languages (e.g., Spanish-English)
- between large resource languages (e.g., French-English)
- in-domain (training and test from the same domain)

→ access to foreign language

SMT could be better

- into morphologically rich / free word order languages (e.g., German)
- handling noisy inputs (e.g., chats, Twitter feeds)
- dealing with documents (instead of sentences)

→ precision / translation accuracy

Conclusion

SMT is a cheap way to access foreign material

Summary

SMT works well

- between similar languages (e.g., Spanish-English)
- between large resource languages (e.g., French-English)
- in-domain (training and test from the same domain)

→ access to foreign language

SMT could be better

- into morphologically rich / free word order languages (e.g., German)
- handling noisy inputs (e.g., chats, Twitter feeds)
- dealing with documents (instead of sentences)

→ precision / translation accuracy

Conclusion

SMT is a cheap way to access foreign material

Summary

SMT works well

- between similar languages (e.g., Spanish-English)
- between large resource languages (e.g., French-English)
- in-domain (training and test from the same domain)

→ access to foreign language

SMT could be better

- into morphologically rich / free word order languages (e.g., German)
- handling noisy inputs (e.g., chats, Twitter feeds)
- dealing with documents (instead of sentences)

→ precision / translation accuracy

Conclusion

SMT is a cheap way to access foreign material

Summary

SMT works well

- between similar languages (e.g., Spanish-English)
- between large resource languages (e.g., French-English)
- in-domain (training and test from the same domain)

→ access to foreign language

SMT could be better

- into morphologically rich / free word order languages (e.g., German)
- handling noisy inputs (e.g., chats, Twitter feeds)
- dealing with documents (instead of sentences)

→ precision / translation accuracy

Conclusion

SMT is a cheap way to access foreign material

Summary

SMT works well

- between similar languages (e.g., Spanish-English)
- between large resource languages (e.g., French-English)
- in-domain (training and test from the same domain)

→ access to foreign language

SMT could be better

- into morphologically rich / free word order languages (e.g., German)
- handling noisy inputs (e.g., chats, Twitter feeds)
- dealing with documents (instead of sentences)

→ precision / translation accuracy

Conclusion

SMT is a cheap way to access foreign material