Statistical Machine Translation

May 13th, 2014

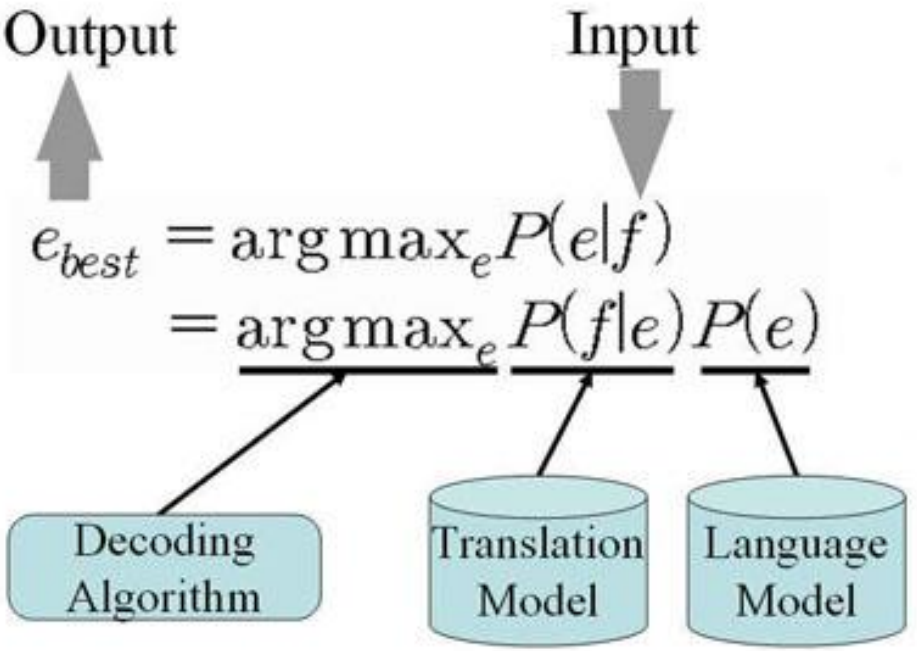
Josef van Genabith
DFKI GmbH
Josef.van_Genabith@dfki.de

Language Technology II
SS 2014

With additional slides from Sabine Hunsiker LT-II SS2012

Overview

- Introduction: the basic idea
- IBM models
- Phrase-Based SMT



Languages and Machines

UNIVERSITÄT DES SAARLANDES

- Human languages are extremely:
  - Elegant
  - Efficient
  - Flexible
- Turns out that this is exactly what makes it difficult for machines to understand human language, translate between them etc.

Languages and Machines

UNIVERSITÄT DES SAARLANDES

What's the matter with Human Language:

- Used to communicate
- record, store, manipulate, share, transmit information …

- Highly ambiguous (a single string can mean many things)
- Flexible (we can say the same thing in many ways)
- Context dependent (meaning changes in context)
- Literal and non-literal meaning, metaphor ….

Languages and Machines

UNIVERSITÄT DES SAARLANDES

…The President's Research Award …

Ambiguity

- The award the president won?
- The award the president is awarding?

Structural

- Who is this President anyway?

Context

- the Áras an Uachtaráin […] … The President's Research Award
- …… the Helix in DCU …… […] … The President's Research Award

- … The President's Research Award ….
- … Prof. Brian MacCraith's Research Award …
- … Brian's Research Award …

Many ways of saying same thing

Languages and Machines

UNIVERSITÄT DES SAARLANDES

- Human Languages:
  - Highly ambiguous (a single string can mean many things)
  - Flexible (we can say the same thing in many ways)
  - Context dependent (meaning changes in context)
  - Literal and non-literal, non-compositional …
- Many-to-many mappings between form ⇔ meaning
- Human Languages ≠ Formal Languages (Math, Logic, …)

Unstructured representations of information!

- Difficult for machines! Humans: context, world knowledge, …

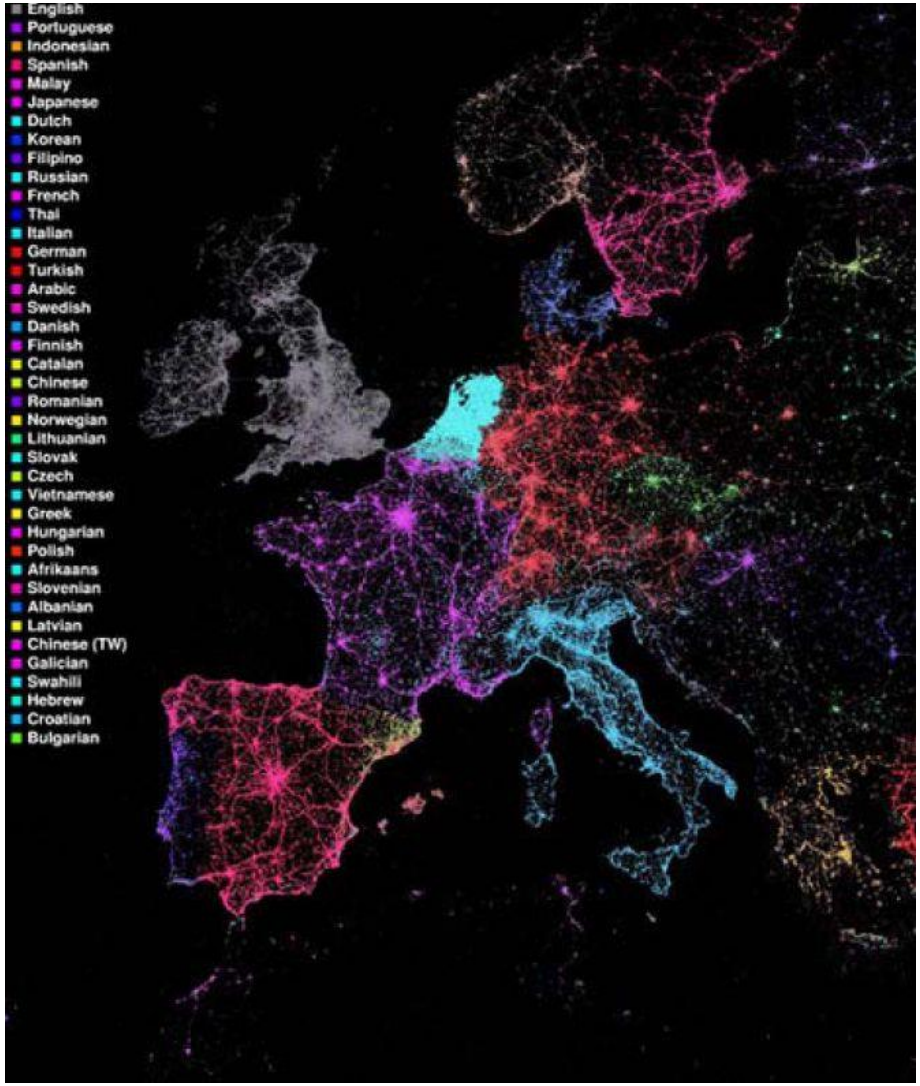
Languages and Machines

UNIVERSITÄT DES SAARLANDES

Human Languages:

- Unstructured representation of information
- Many-to-many mappings between form ⇔ meaning

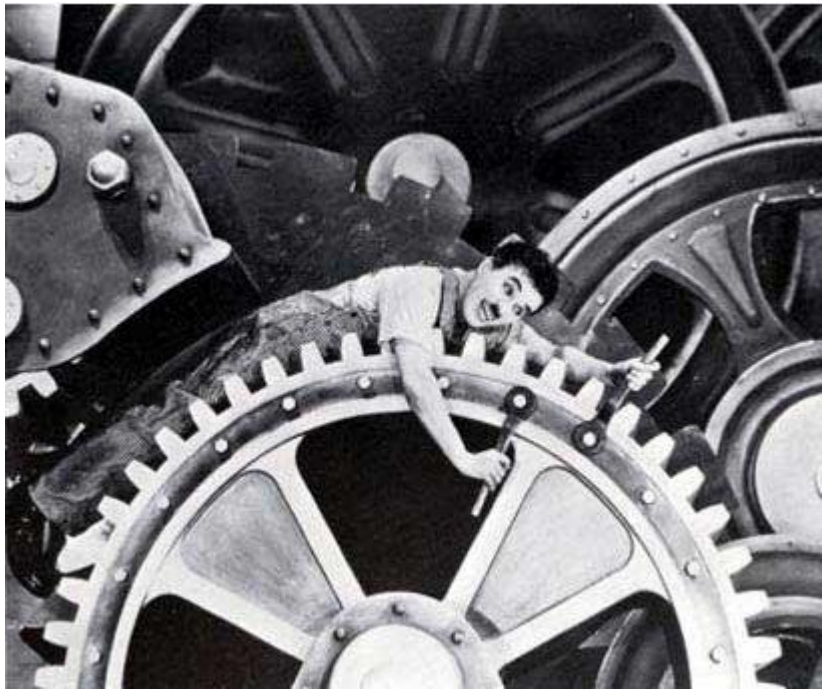
- There are many of them …..!


English
Portuguese
Indonesian
Spanish
Malay
Japanese
Dutch
Korean
Filipino
Russian
French
Thai
Italian
German
Turkish
Arabic
Swedish
Danish
Finnish
Catalan
Chinese
Romanian
Norwegian
Lithuanian
Slovak
Czech
Vietnamese
Greek
Hungarian
Polish
Afrikaans
Slovenian
Albanian
Latvian
Chinese (TW)
Galician
Swahili
Hebrew
Croatian
Bulgarian

Languages and Machines

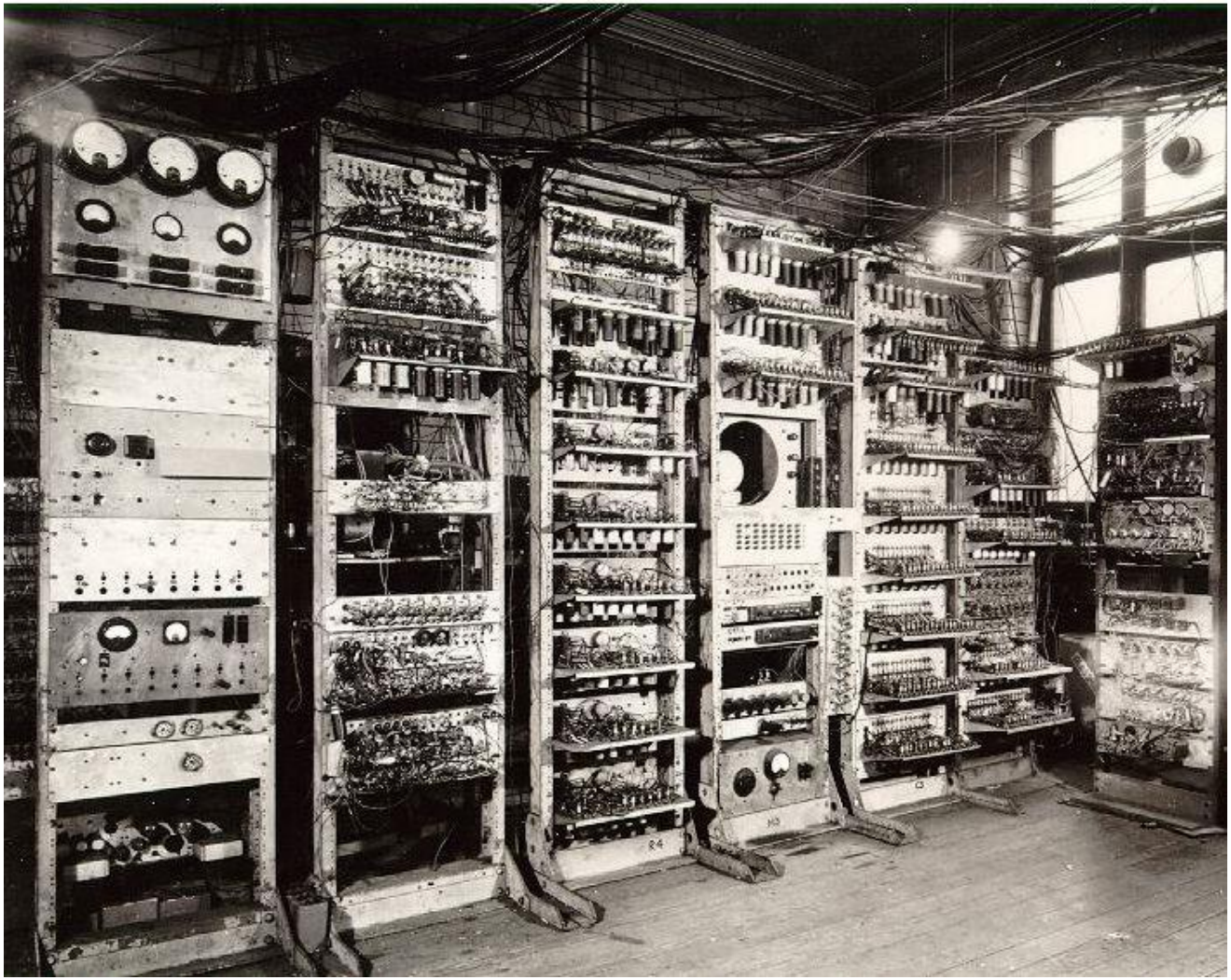
UNIVERSITÄT DES SAARLANDES

- Translation
  - Automatic
  - Fast
  - Scalable
  - Accurate
- Machine Translation MT
- How does MT work?
  - Rule-based MT
  - Statistical MT



Machine Translation

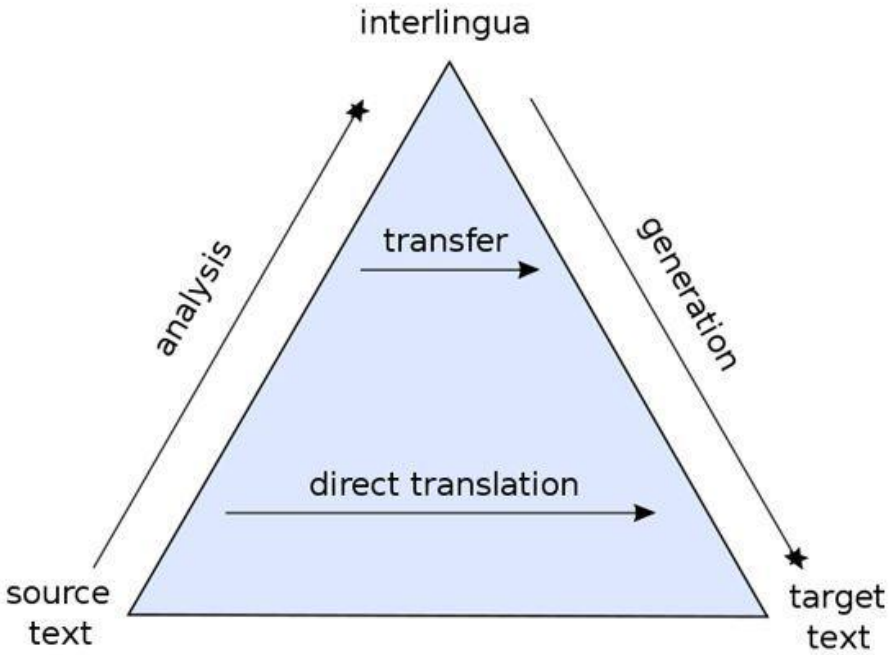
UNIVERSITÄT DES SAARLANDES



Rule-Based Machine Translation

UNIVERSITÄT DES SAARLANDES

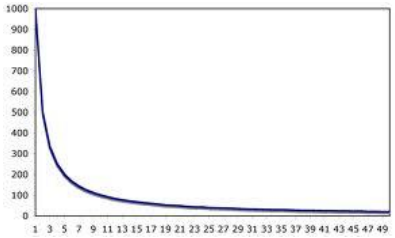
- Rule-Based Machine Translation (RBMT)
- Dominant Paradigm 1950s-1990s
  - Words: Dictionaries
  - Rules: Regularities
  - Hand-crafted
  - Highly skillful knowledge engineer

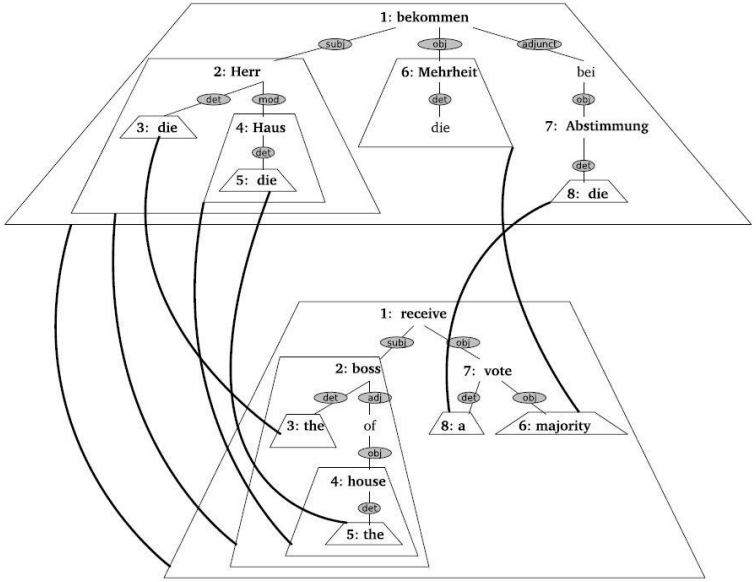


Rule-Based Machine Translation

UNIVERSITÄT DES SAARLANDES

- Problems with RBMT:
  - Languages are complicated
  - Many words
  - Words are ambiguous
  - Lots of rules
  - Lots of exceptions to the rules





  - Very labour intensive to scale
  - 20-100 ++ person years required for each language pair
  - 27 EU languages: > 700 pairs
  - 100 languages: > 9900 pairs

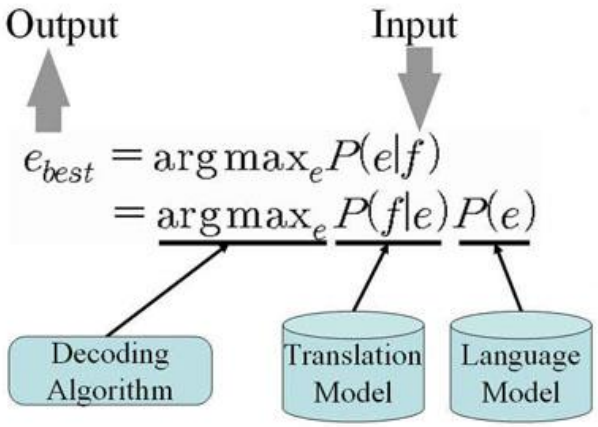
Difficult to scale!

Statistical Machine Translation

UNIVERSITÄT DES SAARLANDES

- Is there a simpler way of doing MT?
- Can machines learn automatically how to translate just from data (= already translated text)?

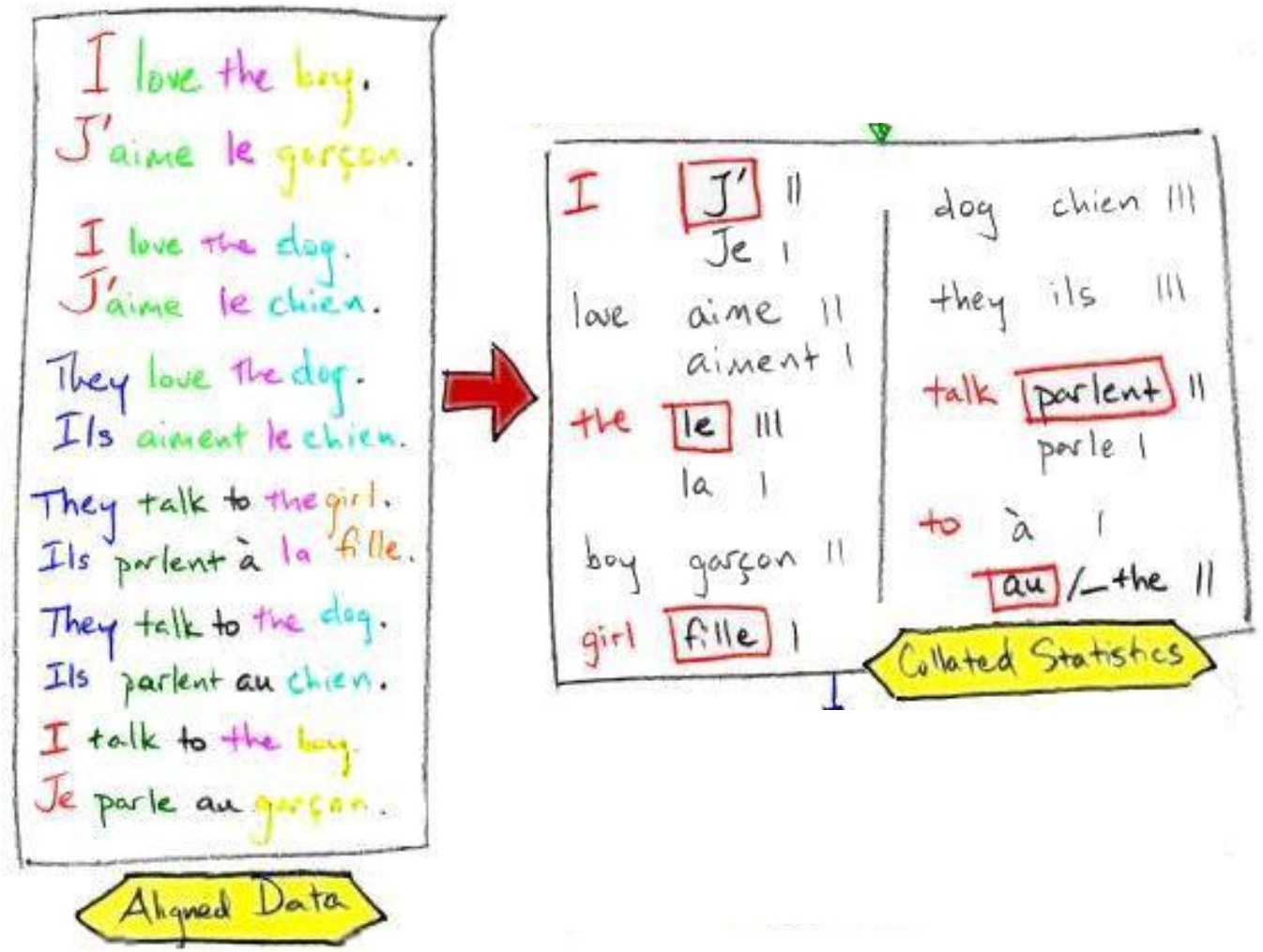
- Data-Driven MT (machine-learning based)
  - Statistical Machine Translation (SMT)



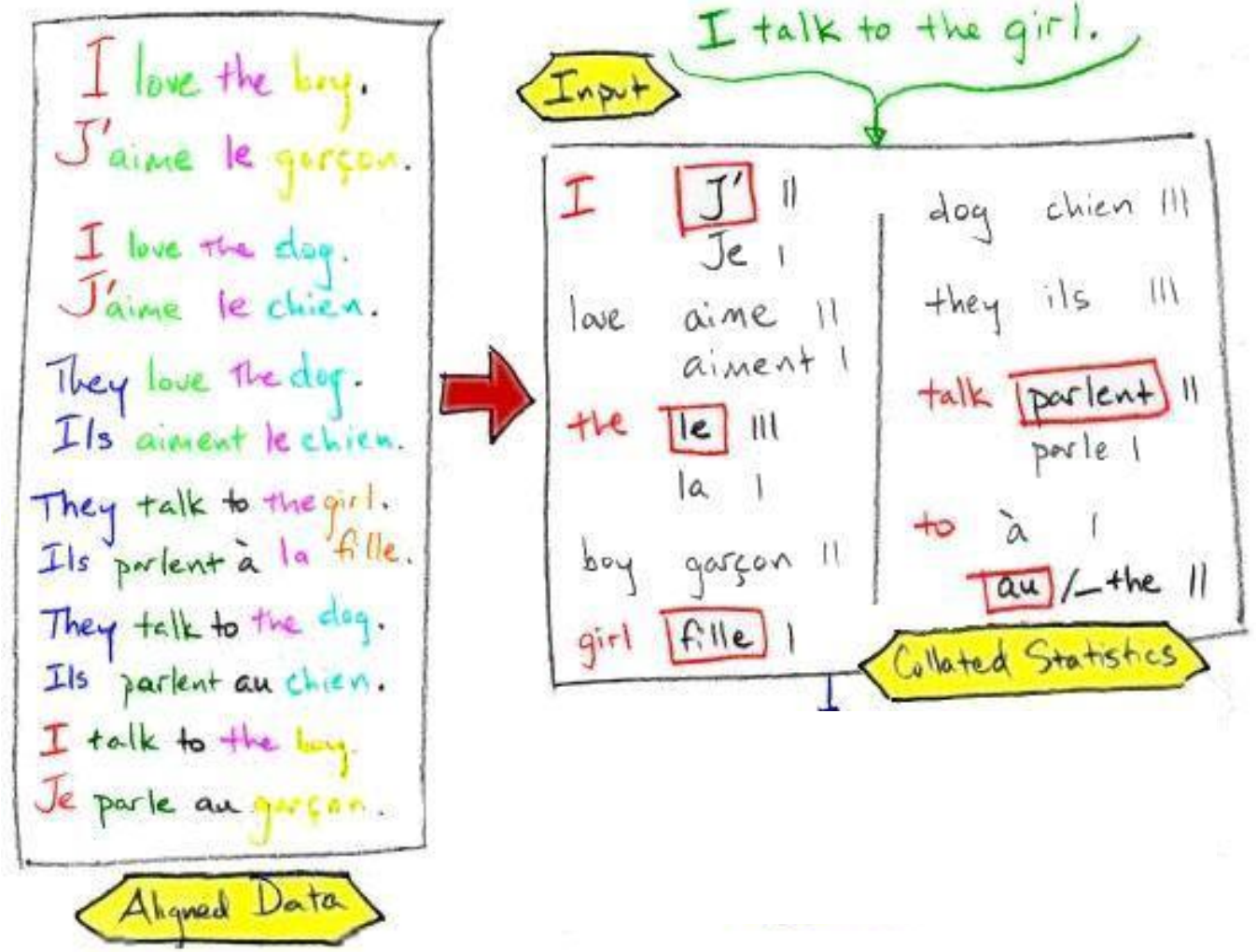
Statistical Machine Translation

UNIVERSITÄT DES SAARLANDES

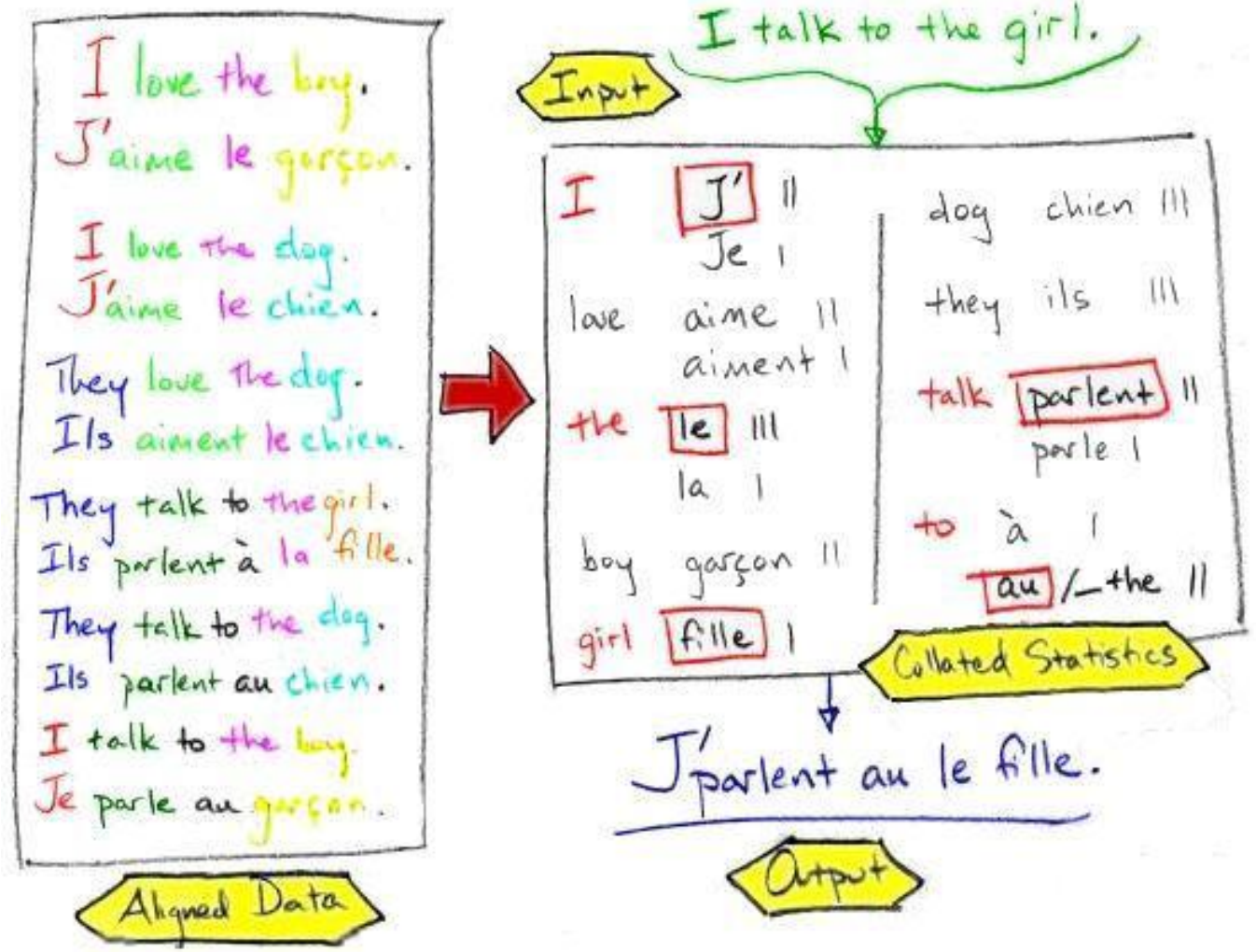
Statistical Machine Translation



Statistical Machine Translation



Statistical Machine Translation



How the computer learns to align/translate words:

UNIVERSITÄT DES SAARLANDES

CLASSIC SOUPS

			Sm.	Lg.
清燉雞湯	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58.	Chicken Rice Soup	1.85	3.25
雞麵湯	59.	Chicken Noodle Soup	1.85	3.25
廣東雲吞	60.	Cantonese Wonton Soup	1.50	2.75
蕃茄蛋湯	61.	Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62.	Regular Wonton Soup	1.10	2.10
酸辣湯	63.	Hot & Sour Soup	1.10	2.10
蛋花湯	64.	Egg Drop Soup	1.10	2.10
雲蛋湯	65.	Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66.	Tofu Vegetable Soup	NA	3.50
雞玉米湯	67.	Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68.	Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69.	Seafood Soup	NA	3.50

How the computer learns to align/translate words:

UNIVERSITÄT DES SAARLANDES

CLASSIC SOUPS

		Sm.	Lg.
清燉雞湯	57. House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58. Chicken Rice Soup	1.85	3.25
雞麵湯	59. Chicken Noodle Soup	1.85	3.25
廣東雲吞	60. Cantonese Wonton Soup	1.50	2.75
蕃茄雪湯	61. Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62. Regular Wonton Soup	1.10	2.10
酸辣湯	63. Hot & Sour Soup	1.10	2.10
雪花湯	64. Egg Drop Soup	1.10	2.10
雲雪湯	65. Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66. Tofu Vegetable Soup	NA	3.50
雞玉米湯	67. Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68. Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69. Seafood Soup	NA	3.50

How the computer learns to align/translate words:

UNIVERSITÄT DES SAARLANDES

CLASSIC SOUPS

			Sm.	Lg.
清燉雞湯	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58.	Chicken Rice Soup	1.85	3.25
雞麵湯	59.	Chicken Noodle Soup	1.85	3.25
廣東雲吞	60.	Cantonese Wonton Soup	1.50	2.75
蕃茄雪湯	61.	Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62.	Regular Wonton Soup	1.10	2.10
酸辣湯	63.	Hot & Sour Soup	1.10	2.10
雪花湯	64.	Egg Drop Soup	1.10	2.10
雲雪湯	65.	Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66.	Tofu Vegetable Soup	NA	3.50
雞玉米湯	67.	Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68.	Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69.	Seafood Soup	NA	3.50

How the computer learns to align/translate words:

UNIVERSITÄT DES SAARLANDES

CLASSIC SOUPS

			Sm.	Lg.
清燉雞湯	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58.	Chicken Rice Soup	1.85	3.25
雞麵湯	59.	Chicken Noodle Soup	1.85	3.25
廣東雲吞	60.	Cantonese Wonton Soup	1.50	2.75
蕃茄雲湯	61.	Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62.	Regular Wonton Soup	1.10	2.10
酸辣湯	63.	Hot & Sour Soup	1.10	2.10
雲花湯	64.	Egg Drop Soup	1.10	2.10
雲雲湯	65.	Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66.	Tofu Vegetable Soup	NA	3.50
雞玉米湯	67.	Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68.	Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69.	Seafood Soup	NA	3.50

How the computer learns to align/translate words:

UNIVERSITÄT DES SAARLANDES

CLASSIC SOUPS

			Sm.	Lg.
清燉雞湯	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58.	Chicken Rice Soup	1.85	3.25
雞麵湯	59.	Chicken Noodle Soup	1.85	3.25
廣東雲吞	60.	Cantonese Wonton Soup	1.50	2.75
蕃茄蛋湯	61.	Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62.	Regular Wonton Soup	1.10	2.10
酸辣湯	63.	Hot & Sour Soup	1.10	2.10
蛋花湯	64.	Egg Drop Soup	1.10	2.10
雲蛋湯	65.	Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66.	Tofu Vegetable Soup	NA	3.50
雞玉米湯	67.	Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68.	Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69.	Seafood Soup	NA	3.50

How the computer learns to align/translate words:

UNIVERSITÄT DES SAARLANDES

CLASSIC SOUPS

		Sm.	Lg.
清燉雞湯	57. House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58. Chicken Rice Soup	1.85	3.25
雞麵湯	59. Chicken Noodle Soup	1.85	3.25
廣東雲吞	60. Cantonese Wonton Soup	1.50	2.75
蕃茄蛋湯	61. Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62. Regular Wonton Soup	1.10	2.10
酸辣湯	63. Hot & Sour Soup	1.10	2.10
蛋湯	64. Egg Drop Soup	1.10	2.10
雲蛋	65. Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66. Tofu Vegetable Soup	NA	3.50
雞玉米湯	67. Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68. Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69. Seafood Soup	NA	3.50

How the computer learns to align/translate words:

UNIVERSITÄT DES SAARLANDES

CLASSIC SOUPS

		Sm.	Lg.
清燉雞湯	57. House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58. Chicken Rice Soup	1.85	3.25
雞麵湯	59. Chicken Noodle Soup	1.85	3.25
廣東雲吞	60. Cantonese Wonton Soup	1.50	2.75
蕃茄蛋湯	61. Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62. Regular Wonton Soup	1.10	2.10
酸辣湯	63. Hot & Sour Soup	1.10	2.10
蛋湯	64. Egg Drop Soup	1.10	2.10
雲蛋湯	65. Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66. Tofu Vegetable Soup	NA	3.50
雞玉米湯	67. Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68. Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69. Seafood Soup	NA	3.50

How the computer learns to align/translate words:

UNIVERSITÄT DES SAARLANDES

CLASSIC SOUPS

		Sm.	Lg.
57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
58.	Chicken Rice Soup	1.85	3.25
59.	Chicken Noodle Soup	1.85	3.25
60.	Cantonese Wonton Soup	1.50	2.75
61.	Tomato Clear Egg Drop Soup	1.65	2.95
62.	Regular Wonton Soup	1.10	2.10
63.	Hot & Sour Soup	1.10	2.10
64.	Egg Drop Soup	1.10	2.10
65.	Egg Drop Wonton Mix	1.10	2.10
66.	Tofu Vegetable Soup	NA	3.50
67.	Chicken Corn Cream Soup	NA	3.50
68.	Crab Meat Corn Cream Soup	NA	3.50
69.	Seafood Soup	NA	3.50

How the computer learns to align/translate words:

UNIVERSITÄT DES SAARLANDES

CLASSIC SOUPS

			Sm.	Lg.
清燉雞	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯	58.	Chicken Rice Soup	1.85	3.25
雞麵	59.	Chicken Noodle Soup	1.85	3.25
廣東雲吞	60.	Cantonese Wonton Soup	1.50	2.75
蕃茄蛋	61.	Tomato Clear Egg Drop Soup	1.65	2.95
雲吞	62.	Regular Wonton Soup	1.10	2.10
酸辣	63.	Hot & Sour Soup	1.10	2.10
蛋	64.	Egg Drop Soup	1.10	2.10
雲蛋	65.	Egg Drop Wonton Mix	1.10	2.10
豆腐菜	66.	Tofu Vegetable Soup	NA	3.50
雞玉米	67.	Chicken Corn Cream Soup	NA	3.50
蟹肉玉米	68.	Crab Meat Corn Cream Soup	NA	3.50
海鮮	69.	Seafood Soup	NA	3.50

How the computer learns to align/translate words:

UNIVERSITÄT DES SAARLANDES

CLASSIC SOUPS

			Sm.	Lg.
清燉雞	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯	58.	Chicken Rice Soup	1.85	3.25
雞麵	59.	Chicken Noodle Soup	1.85	3.25
廣東雲吞	60.	Cantonese Wonton Soup	1.50	2.75
蕃茄蛋	61.	Tomato Clear Egg Drop Soup	1.65	2.95
雲吞	62.	Regular Wonton Soup	1.10	2.10
酸辣	63.	Hot & Sour Soup	1.10	2.10
蛋	64.	Egg Drop Soup	1.10	2.10
雲蛋	65.	Egg Drop Wonton Mix	1.10	2.10
豆腐菜	66.	Tofu Vegetable Soup	NA	3.50
雞玉米	67.	Chicken Corn Cream Soup	NA	3.50
蟹肉玉米	68.	Crab Meat Corn Cream Soup	NA	3.50
海鮮	69.	Seafood Soup	NA	3.50

How the computer learns to align/translate words:

UNIVERSITÄT DES SAARLANDES

CLASSIC SOUPS

		Sm.	Lg.
清燉雞湯	57. House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58. Chicken Rice Soup	1.85	3.25
雞麵湯	59. Chicken Noodle Soup	1.85	3.25
廣東雲吞	60. Cantonese Wonton Soup	1.50	2.75
蕃茄雪湯	61. Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62. Regular Wonton Soup	1.10	2.10
酸辣湯	63. Hot & Sour Soup	1.10	2.10
雪花湯	64. Egg Drop Soup	1.10	2.10
雲雪湯	65. Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66. Tofu Vegetable Soup	NA	3.50
雞玉米湯	67. Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68. Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69. Seafood Soup	NA	3.50

How the computer learns to align/translate words:

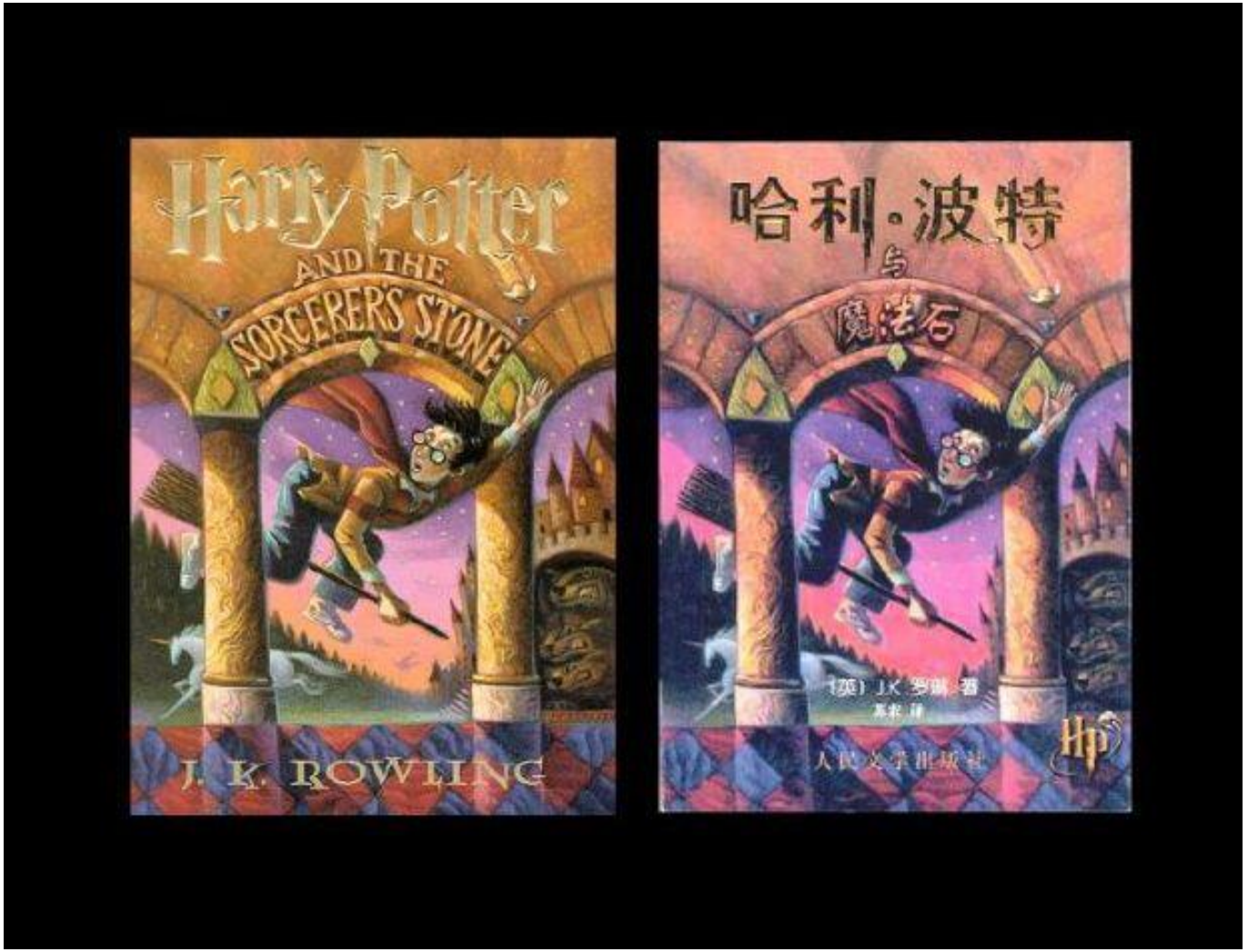
UNIVERSITÄT DES SAARLANDES

CLASSIC SOUPS

		Sm.	Lg.
清燉雞湯	57. House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58. Chicken Rice Soup	1.85	3.25
雞麵湯	59. Chicken Noodle Soup	1.85	3.25
廣東雲吞	60. Cantonese Wonton Soup	1.50	2.75
蕃茄雲湯	61. Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62. Regular Wonton Soup	1.10	2.10
酸辣湯	63. Hot & Sour Soup	1.10	2.10
雲花湯	64. Egg Drop Soup	1.10	2.10
雲雲湯	65. Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66. Tofu Vegetable Soup	NA	3.50
雞玉米湯	67. Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68. Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69. Seafood Soup	NA	3.50

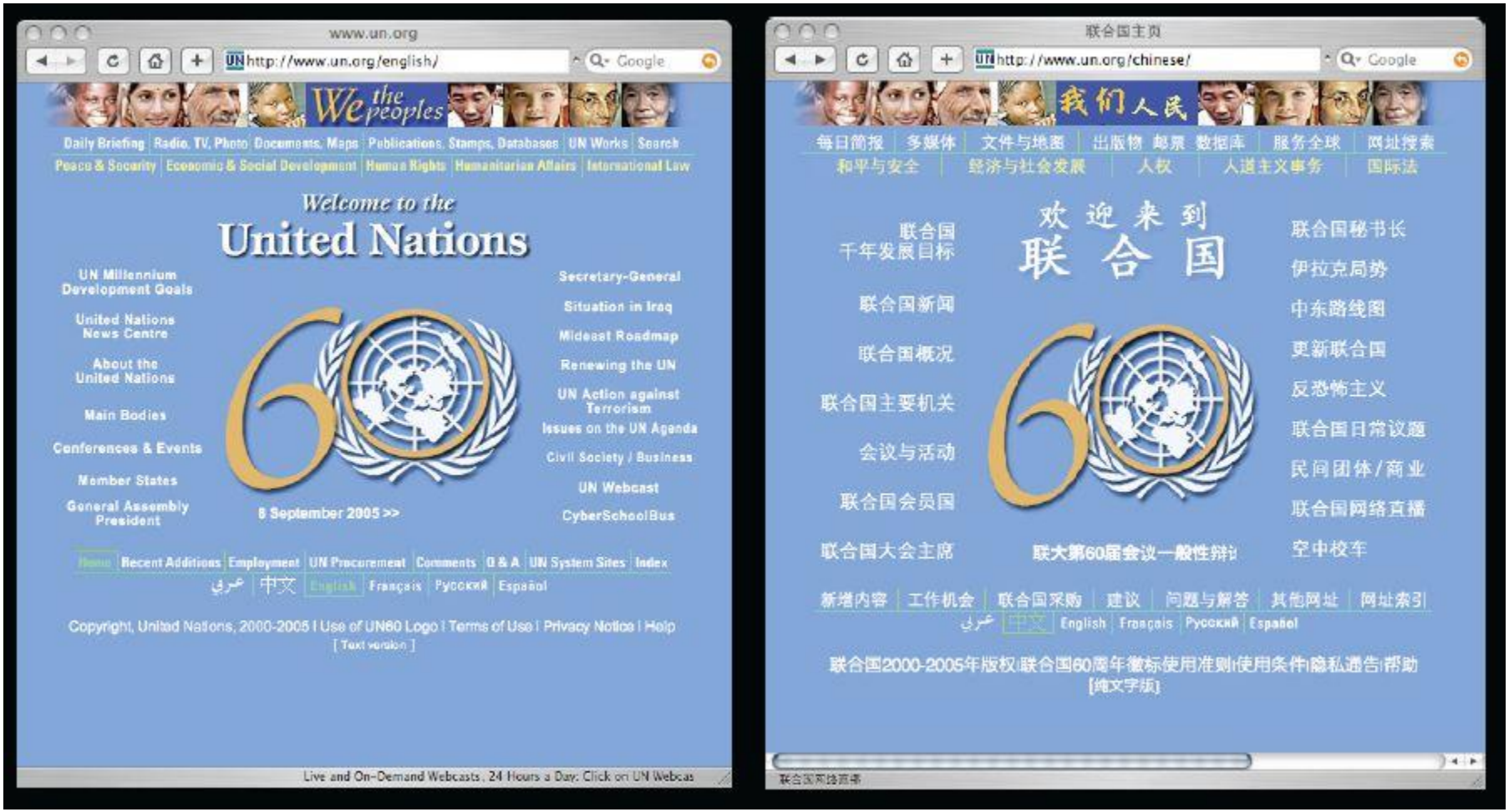
Data

UNIVERSITÄT DES SAARLANDES



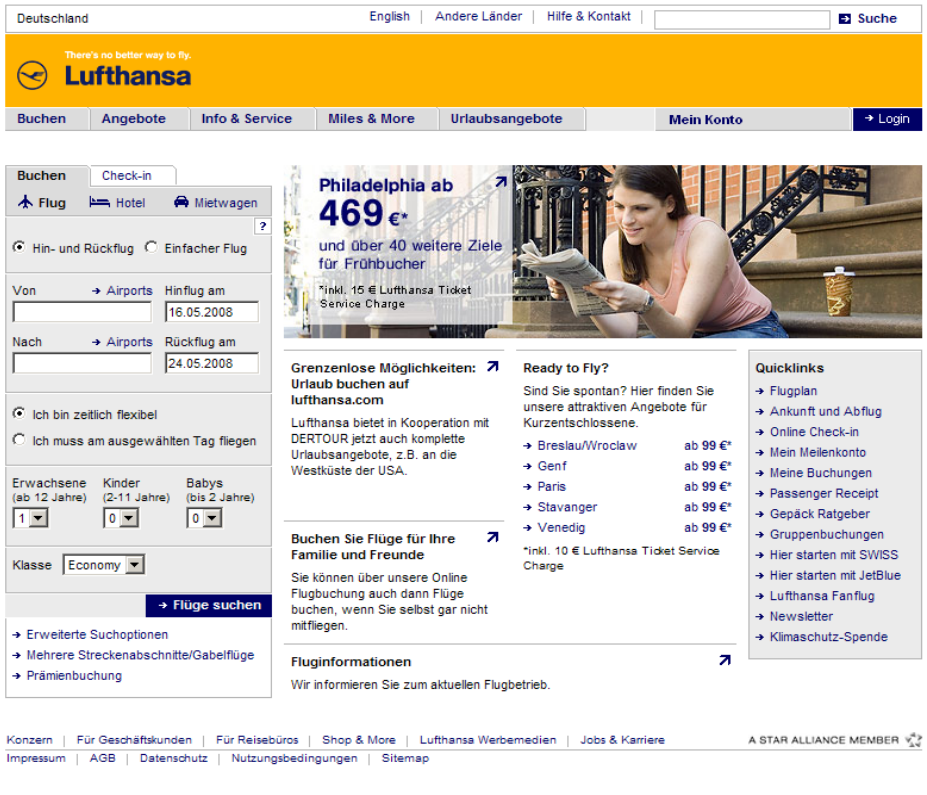
Data

UNIVERSITÄT DES SAARLANDES



Data

UNIVERSITÄT DES SAARLANDES

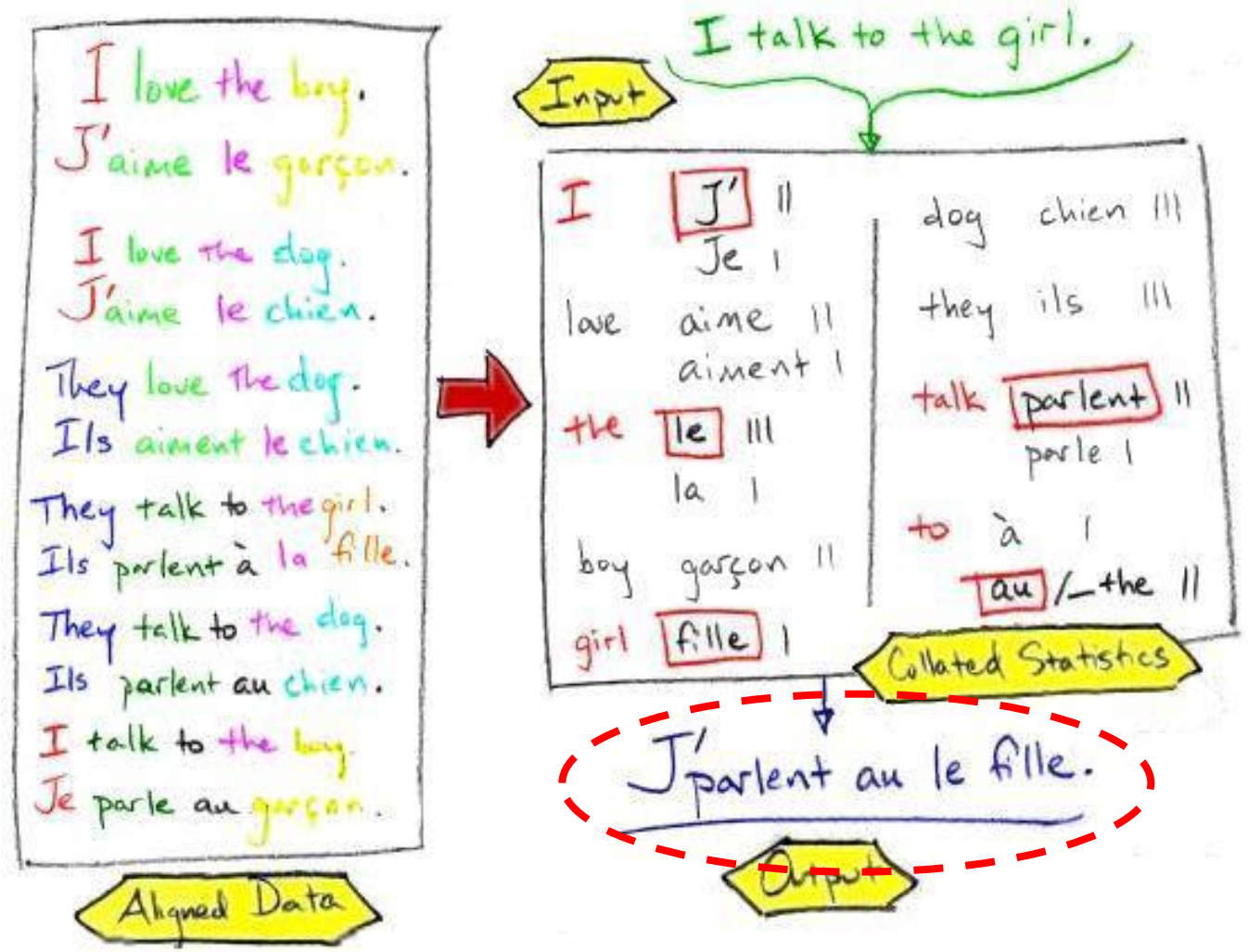


Data

UNIVERSITÄT DES SAARLANDES

- Data
  - Translations (electronic form)
  - Translation Memories
  - WWW
- Europarl: translations of European Parliamentary Debates
- French-English corpus: 100M words

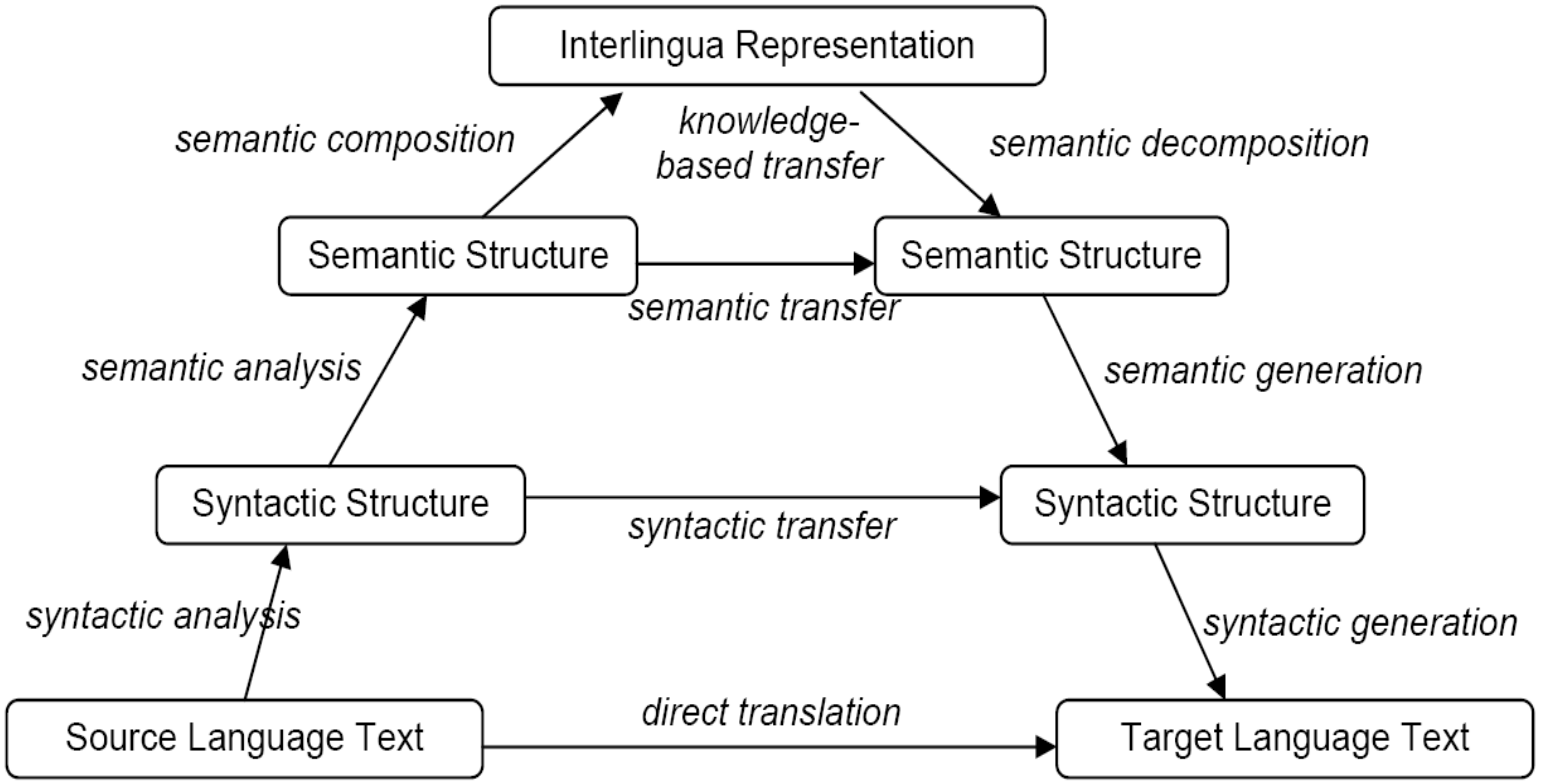
Statistical Machine Translation



Possible (rule-based) MT architectures

UNIVERSITÄT DES SAARLANDES

The „Vauquois Triangle“



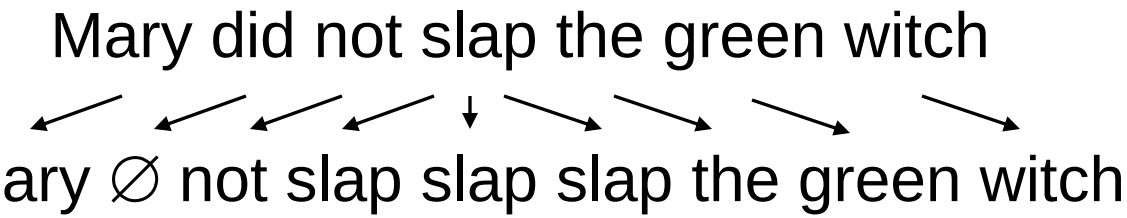
What do we need to model?

UNIVERSITÄT DES SAARLANDES

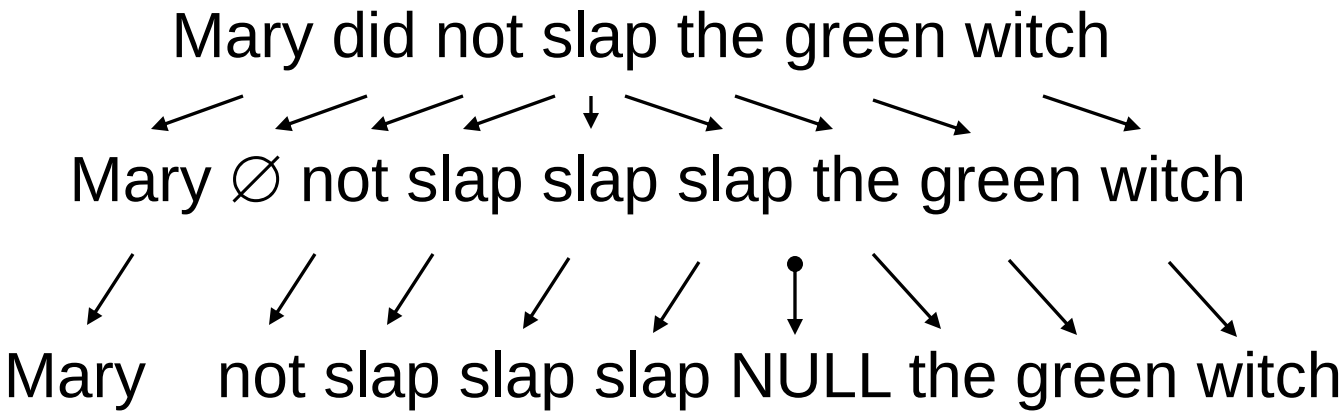
Mary did not slap the green witch

Maria no daba una bofetada a la bruja verde

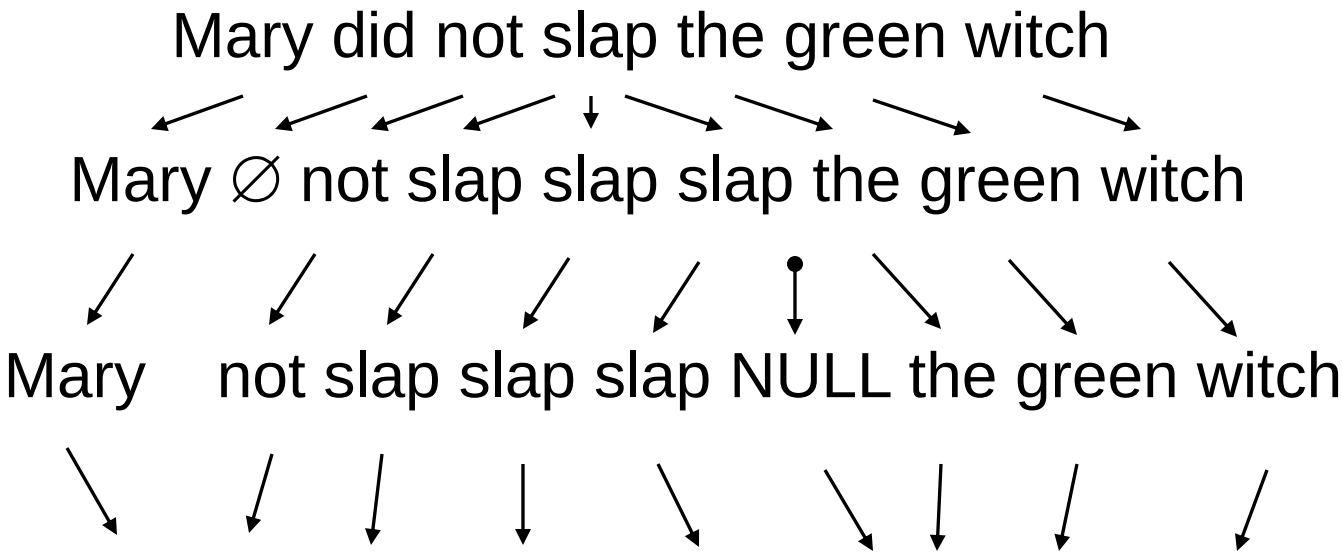
What do we need to model?



Maria no daba una bofetada a la bruja verde

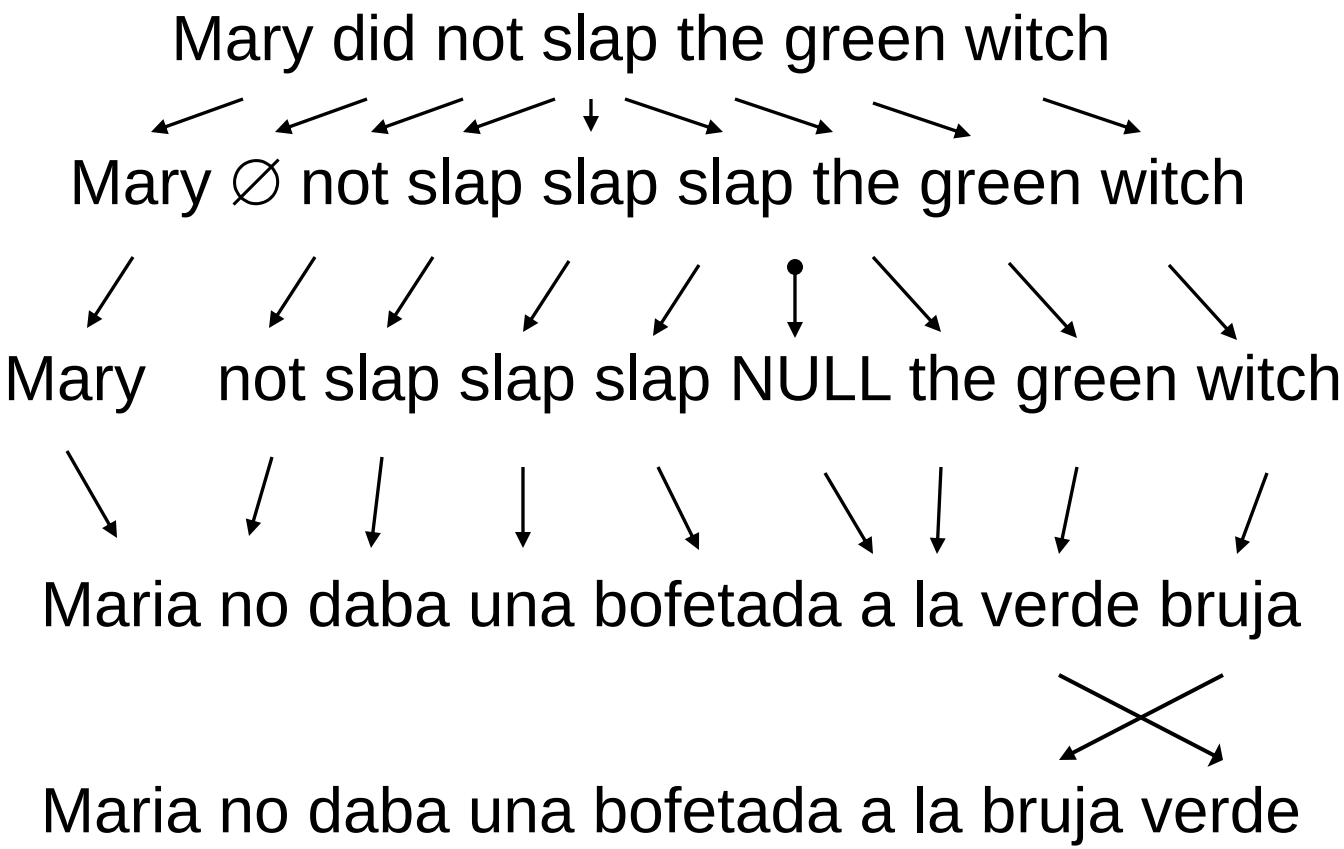


Maria no daba una bofetada a la bruja verde



Maria no daba una bofetada a la bruja verde

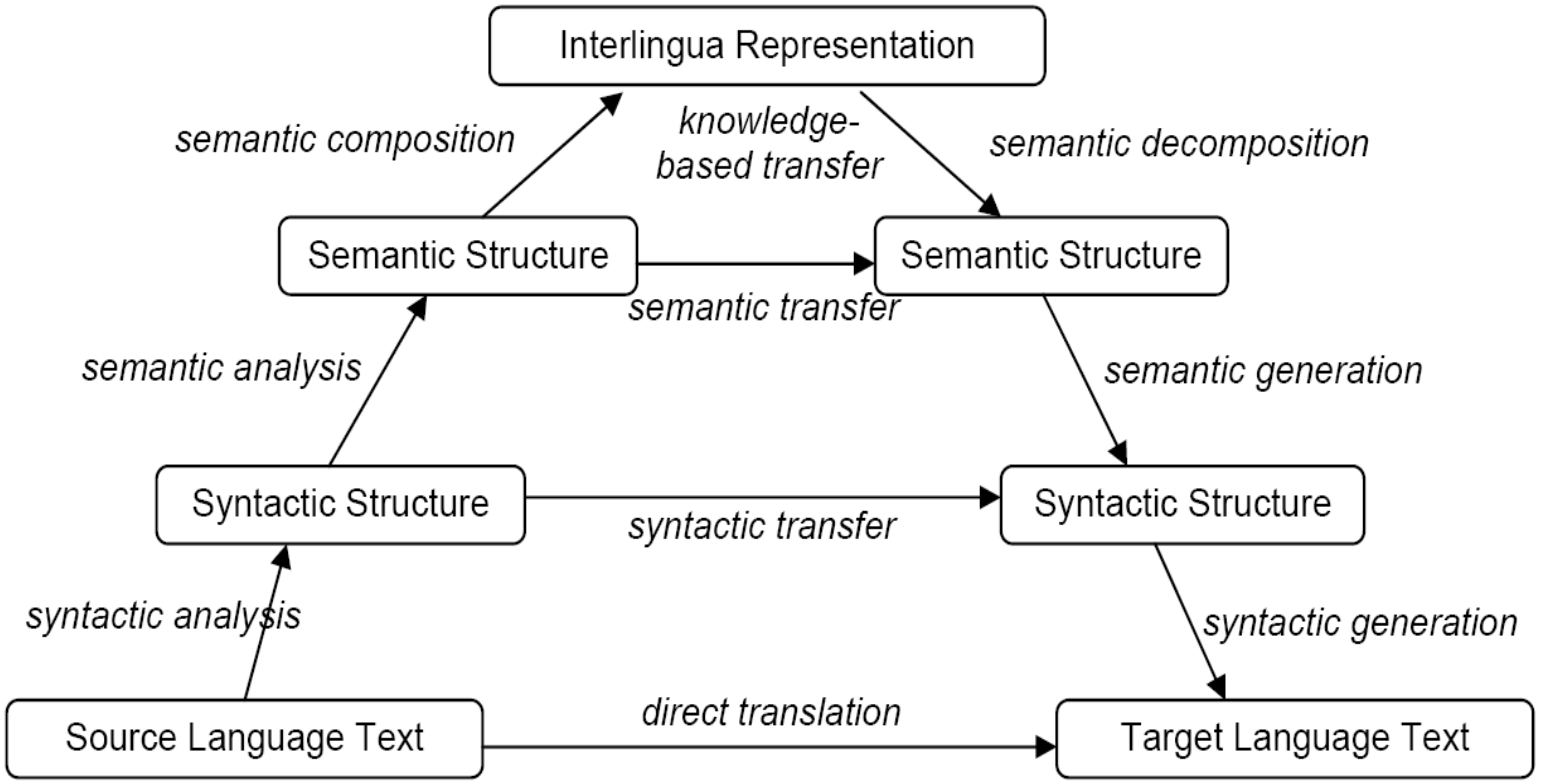
What do we need to model?



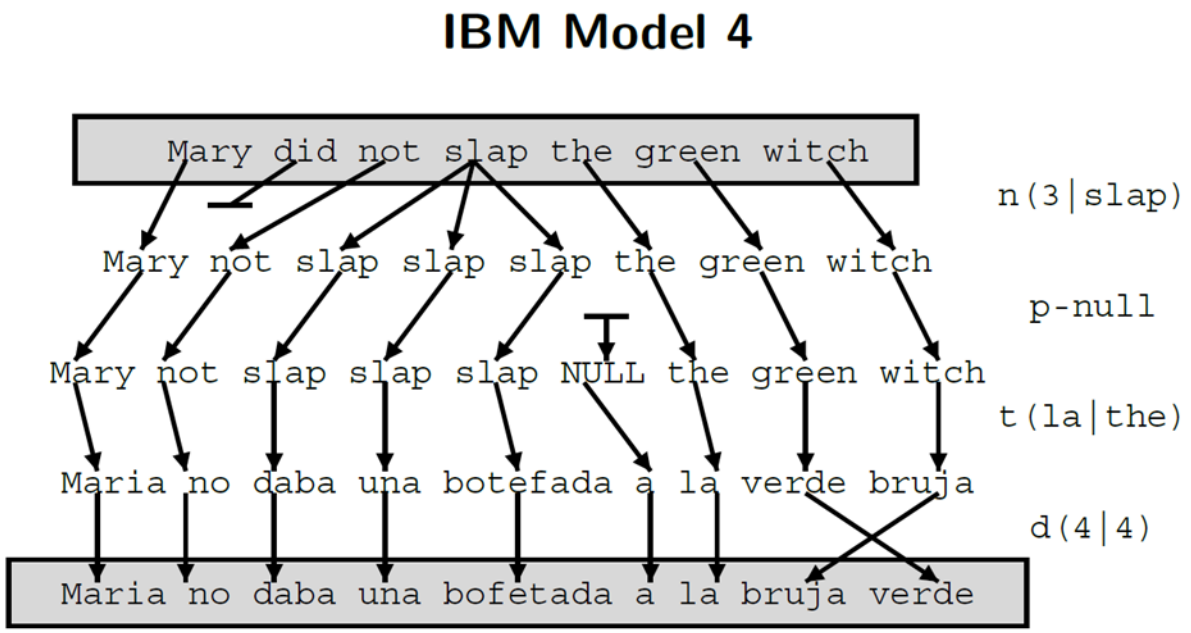
Possible (rule-based) MT architectures

The „Vauquois Triangle“

Slide by Sabine Hunsiker



What do we need to model?



More Slides by Sabine Hunsiker SS2012

UNIVERSITÄT DES SAARLANDES

Slides from the SS2012 version of the course

Sources for Information

UNIVERSITÄT DES SAARLANDES

- MT in general, history:
  - http://www.MT-Archive.info: Electronic repository and bibliography of articles, books and papers on topics in machine translation and computer-based translation tools, regularly updated, contains over 3300 items
  - Hutchins & Somers: An introduction to machine translation. Academic Press, 1992, available under http://www.hutchinsweb.me.uk/IntroMT-TOC.htm
- MT systems:

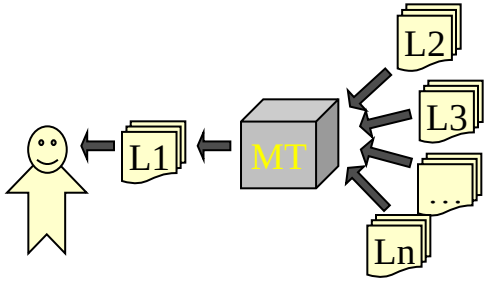
  Compendium of Translation Software, see http://www.hutchinsweb.me.uk/Compendium.htm
- Statistical Machine Translation:

  See www.statmt.org

  Book by Philipp Koehn is available in the coli-bib

Use cases and requirements for MT

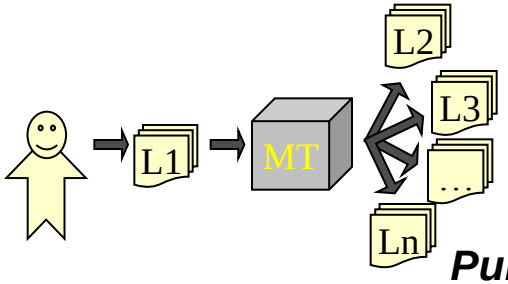
a) MT for assimilation „inbound“



Robustness
Coverage

Daily throughput of online-MT-Systems > 500 M Words

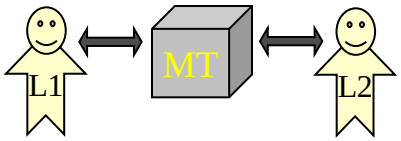
b) MT for dissemination „outbound“



Textual quality

Publishable quality can only be authored by humans; Translation Memories & CAT-Tools mandatory for professional translators

c) MT for direct communication



Speech recognition, context dependence

Topic of many running and completed research projects (VerbMobil, TC Star, TransTac, …)
US-Military uses systems for spoken MT

On the Risks of Outbound MT

UNIVERSITÄT DES SAARLANDES

Some recent examples

'I am not in the office at the moment. Please send any work to be translated'

Motivation for statistical MT

UNIVERSITÄT DES SAARLANDES

- Good translation requires knowledge and decisions on many levels
  - syntactic disambiguation (POS, attachments)
  - semantic disambiguation (collocations, scope, word sense)
  - reference resolution
  - lexical choice in target language
  - application-specific terminology, register, connotations, good style …
- Rule-based models of all these levels are very expensive to build, maintain, and adapt to new domains
- Statistical approaches have been quite successful in many areas of NLP, once data has been annotated
- Learning from existing translation will focus on distinctions that matter (not on the linguist's favorite subject)
- Translation corpora are available in rapidly growing amounts
- SMT *can* integrate rule-based modules (morphologies, lexicons)
- SMT *can* use feed-back for on-line adaptation to domain and user preferences

History of SMT and Important Players I

UNIVERSITÄT DES SAARLANDES

- 1949: Warren Weaver: *the translation problem can be largely solved by "statistical semantic studies"*
- 1950s..1970s: Predominance of rule-based approaches
- 1966: ALPAC report: general discouragement for MT (in the US)
- 1980s: example-based MT proposed in Japan (Nagao), statistical approaches to speech recognition (Jelinek et al. at IBM)
- Late 80s: Statistical POS taggers, SMT models at IBM, work on translation alignment at Xerox (M. Kay)
- Early 90s: many statistical approaches to NLP in general, IBM's Candide claimed to be as good as Systran
- Late 90s: Statistical MT successful as a fallback approach within Verbmobil System (Ney, Och). Wide distribution of translation memory technology (Trados) indicates big commercial potential of SMT
- 1999 Johns Hopkins workshop: open source re-implementation of IBM's SMT methods (GIZA)

History of SMT and Important Players II

UNIVERSITÄT DES SAARLANDES

- Since 2001: DARPA/NIST evaluation campaign (XYZ → English), uses BLEU score for automatic evaluation
- Various companies start marketing/exploring SMT:
  language weaver, aixplain GmbH, Linear B Ltd., esteam, Google Labs
- 2002: Philipp Koehn (ISI) makes EuroParl corpus available
- 2003: Koehn, Och & Marcu propose *Statistical Phrase-Based MT*
- 2004: ISI publishes Philipp Koehn's SMT decoder *Pharaoh*
- 2005: First SMT workshop with shared task
- 2006: Johns Hopkins workshop on OS factored SMT decoder Moses, Start of EuroMatrix project for MT between all EU languages, Acquis Communautaire (EU laws in 20+ languages) made available
- 2007: Google abandons Systran and switches to own SMT technology
- 2009: Start of EuroMatrixPlus *"bringing MT to the user"*
- 2010: Start of many additional MT-related EU projects (Let's MT, ACCURAT, …)