

The concept and feasibility of modern statistical machine translation

Theory, practice and industry application

Maxim Khalilov
TAUS Labs
Amsterdam

Marta R. Costa-jussà
Barcelona Media
Barcelona



RuSSIR 2012
August 5-10



Introducing ourselves and the course

MT is commercially and academically interesting

Commercially	Academically
US has invested in MT for intelligence purposes	MT requires or benefits from many other NLP technologies
MT is popular on the web, it is the most used of Google's special features	Parsing, generation, word sense disambiguation, named entity recognition, transliteration, pronoun resolution, real-world knowledge...
EU spends more than 1,000,000,000 euros on translation costs each year	MT is not going to be solved in the next years

3/5



Introducing ourselves and the course

Who are we?

- Dr Maxim Khalilov -- Research and development manager at TAUS B.V. (Amsterdam). Internship at Macquarie University (Australia), post-doctoral researcher at UvA (Amsterdam).
E-mail: maxim@tauslabs.com
- Dr Marta R. Costa-jussà -- Research fellow at Barcelona Media Innovation Center (Barcelona). Visiting researcher at LIMSI-CNRS (Paris) and I2R (Singapore) and visiting professor at IME-USP (Sao Paulo).
E-mail: marta.ruiz@barcelonamedia.org

Both did his/her PhD at Universitat Politècnica de Catalunya

2/5



Introducing ourselves and the course

Course overview

Day	Content
1	- Historical overview, basic MT concepts - Theory of SMT, state-of-the-art of the technology. Part 1.
2	- Theory of SMT, state-of-the-art of the technology. Part 2. - Evaluation of translation quality.
3	- Key problems in SMT. - Recent SMT approaches
4	- Practical workshop
5	- Seminar - Industry perspectives - Example of SMT systems in production

4/5



Introducing ourselves and the course

What will you learn in this course?

- ▶ We will give you an introduction to the state-of-the-art of the SMT technology.
- ▶ We will provide you with an overview of current problems that MT community is facing.
- ▶ You will gain in-hand experience building SMT engines yourselves from scratch to the quality evaluation step.
- ▶ We will have a good time!

Outline

- ▶ History of MT
- ▶ MT approaches
 - ▶ rule-based and corpus-based (EBMT and SMT)
- ▶ SMT approaches
 - ▶ phrase, hierarchical and syntax-based
- ▶ SMT evaluation
- ▶ SMT scientific community
 - ▶ Challenges, Resources, Software, Conferences/journals, Evaluation Campaigns

▶ 5/5

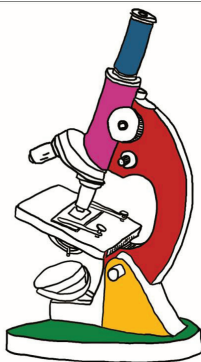


The concept and feasibility of modern statistical machine translation

Historical overview and basic concepts

Maxim Khalilov
TAUS Labs
Amsterdam

Marta R. Costa-jussà
Barcelona Media
Barcelona



RuSSIR 2012
August 5-10



▶ 2

of 31

Georgetown-IBM Experiment

New York, January 7, 1954: Russian was translated into English by an electronic "brain" today for the first time. [...]



▶ 3

of 31

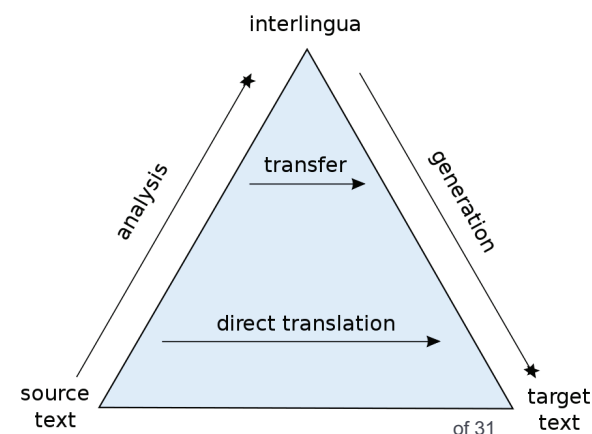
MT, one of the first applications envisioned for computers

- ▶ 1947-1954 Information Theory Foundations
- ▶ 1954-1966 Large bilingual dictionaries + Rules
- ▶ 1966-1980 ALPAC Report, research in Europe, Canada and Soviet Union
- ▶ 1980s Variety of systems
- ▶ 1990s Statistical Machine Translation
- ▶ 2000s MT software

▶ 4

of 31

Vauquois pyramid, comparative depths of MT intermediary representation



▶ 6

of 31

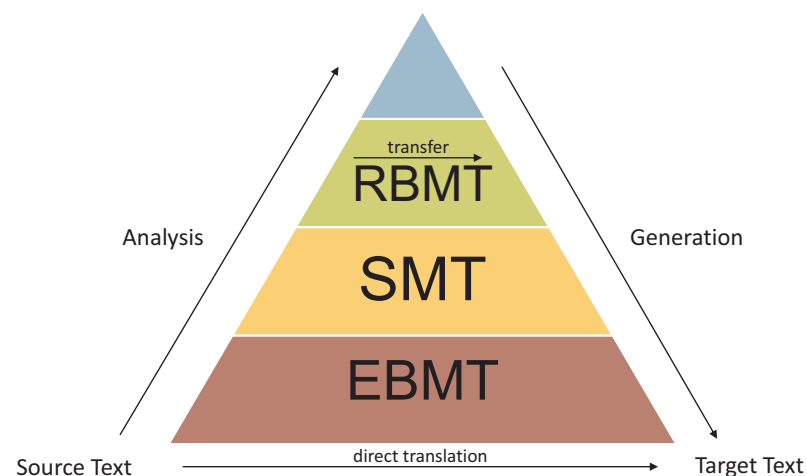
Outline

- ▶ History of MT
- ▶ MT approaches
 - ▶ rule-based and corpus-based (EBMT and SMT)
- ▶ SMT approaches
 - ▶ phrase, hierarchical and syntax-based
- ▶ SMT evaluation
- ▶ SMT scientific community
 - ▶ Challenges, Resources, Software, Conferences/journals, Evaluation Campaigns

▶ 5

of 31

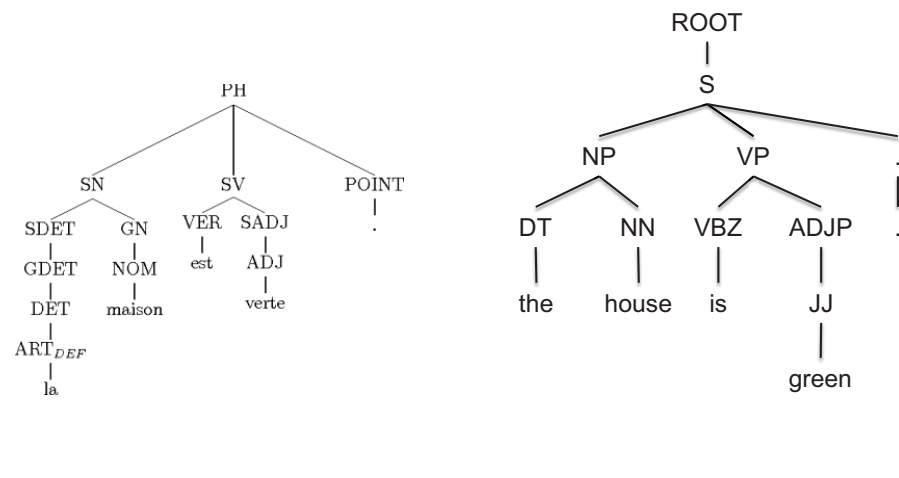
Approaches to Machine Translation



▶ 7

of 31

Rule-based Machine Translation

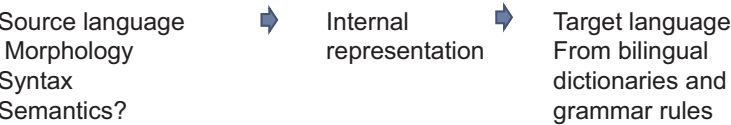


Rule-based MT has several depths of intermediary representation

Interlingua:



Transfer:



What is a Parallel Text or Parallel Corpus?

- Translated text/documents in two languages
- Ideally sentence-aligned

Table 2
Output from alignment program.

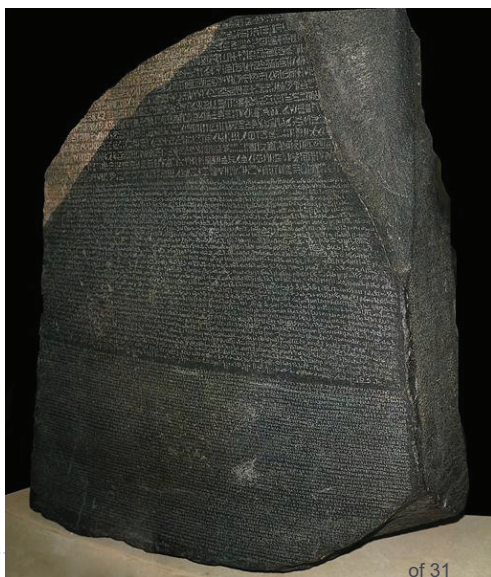
English	French
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates.	Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.
The higher turnover was largely due to an increase in the sales volume.	La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes.
Employment and investment levels also climbed.	L'emploi et les investissements ont également augmenté.

Corpus-based MT are trained on parallel corpora

Collections of parallel texts at sentence level

English	Russian
This course is a thorough introduction to machine translation technology	Этот курс представляет собой интенсивное введение в технологию машинного перевода
We will describe all aspects of building a statistical machine translation system, from both formal and practical perspectives.	Мы рассмотрим все аспекты построения системы статистического машинного перевода с теоретической и практической точки зрения

An early parallel text

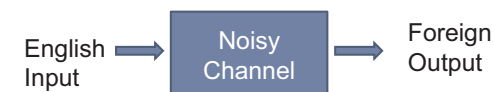


12

of 31

Corpus-based MT has no depth of intermediary representation

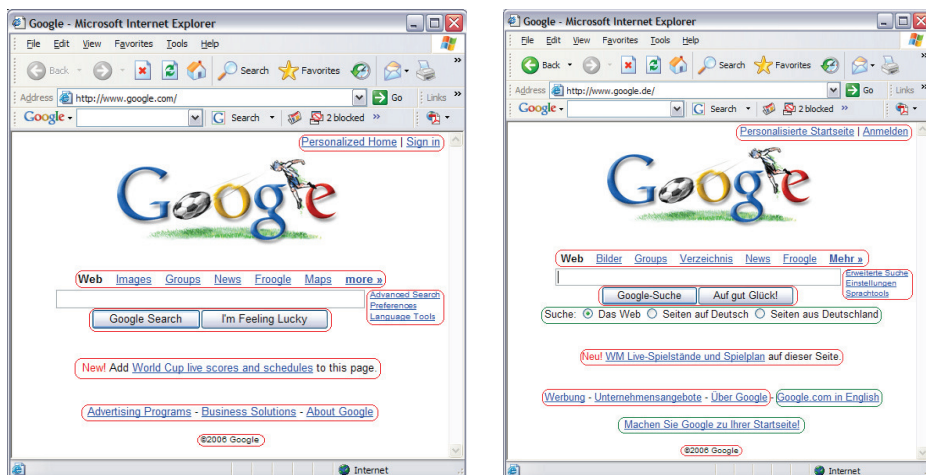
- ▶ Example-based:
translation by analogy
- ▶ Statistical-based:
translation generated on the basis of statistical models



14

of 31

Parallel Text on the Web



13

of 31

Advantages of SMT

- ▶ Data driven
- ▶ Language independent
- ▶ No need for staff of linguists or language experts
- ▶ Can prototype a new system quickly and at a very low cost
- ▶ High flexibility of matching heuristics

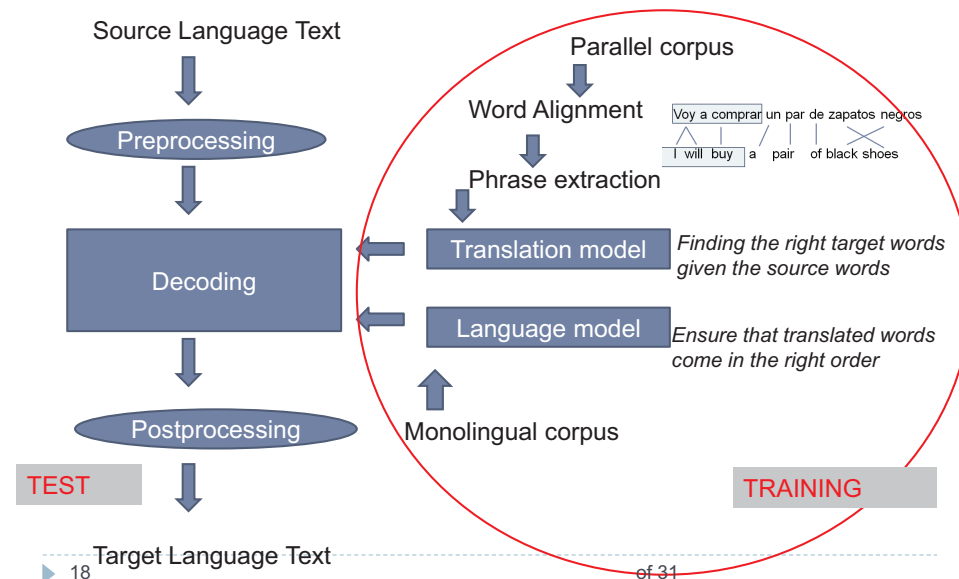
15

of 31

Outline

- ▶ History of MT
- ▶ MT approaches
 - ▶ rule-based and corpus-based (EBMT and SMT)
- ▶ SMT approaches
 - ▶ phrase, hierarchical and syntax-based
- ▶ SMT evaluation
- ▶ SMT scientific community
 - ▶ Challenges, Resources, Software, Conferences/journals, Evaluation Campaigns

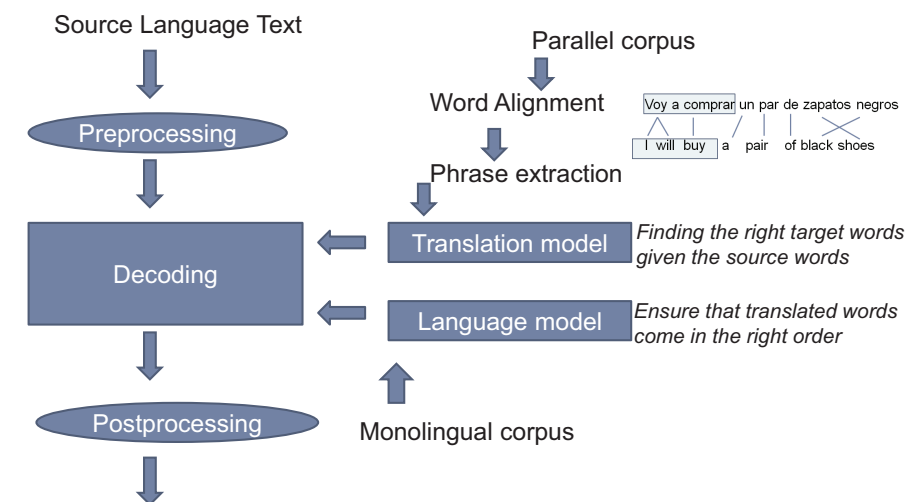
A picture is worth a million equations



▶ 16

of 31

Phrase-based deals with sequences of words



▶ 17

of 31

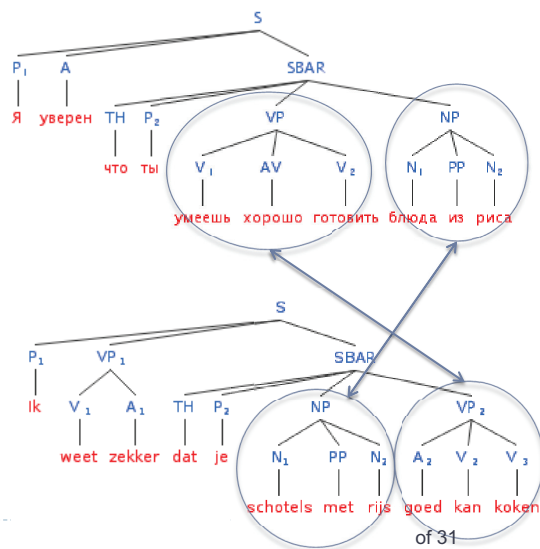
Hierarchical-based introduce hierarchical rules in decoding

- ▶ Hierarchical rules allow for hierarchical phrases that can contain other phrases
- ▶ [Я] [уверен] [что] [ты] [хорошо умеешь готовить] [блюда с рисом]
- ▶ [Ik] [weet zeker] [dat] [je] [schotels met rijst] [goed kan koken]
- ▶ [ты][X][блюда с рисом] - > [je][schotels met rijst][X]

▶ 19

of 31

Syntax Augmented introduce syntax trees in decoding



Outline

- ▶ History of MT
- ▶ MT approaches
 - ▶ rule-based and corpus-based (EBMT and SMT)
- ▶ SMT approaches
 - ▶ phrase, hierarchical and syntax-based
- ▶ SMT evaluation
- ▶ SMT scientific community
 - ▶ Challenges, Resources, Software, Conferences/journals, Evaluation Campaigns

▶ 20

▶ 22

of 31

Example-Based Machine Translation

- ▶ Simplest case
 - ▶ Sentence to be translated matches previously seen sentence
 - ▶ Same as 100% translation memory match
- ▶ Pattern recognition

English	Japanese
How much is that red umbrella ?	Ano akai kasa wa ikura desu ka.
How much is that small camera ?	Ano chiisai kamera wa ikura desu ka.

- ▶ Automatic evaluation allows to optimize the systems.
 - ▶ BLEU is the standard measure taken by the scientific community. It evaluates sequences of ngrams
- ▶ Human evaluation allows for a fair comparison across systems
 - ▶ FLUENCY, ACCURACY are the two standard measures

▶ 21

of 31

▶ 23

of 31

Outline

- ▶ History of MT
- ▶ MT approaches
 - ▶ rule-based and corpus-based (EBMT and SMT)
- ▶ SMT approaches
 - ▶ phrase, hierarchical and syntax-based
- ▶ SMT evaluation
- ▶ SMT scientific community
 - ▶ Challenges, Resources, Software, Conferences/journals, Evaluation Campaigns

The parallel corpus is the main required resource for SMT

- ▶ Parallel corpus:
 - ▶ EPPS,
 - ▶ JRC-Acquis,
 - ▶ UN data
 - ▶ Canadian Hansards
 - ▶ Hong Kong laws parallel text
- ...
- <http://www ldc.upenn.edu>

▶ 24

of 31

SMT challenges are found in all linguistic areas

- ▶ Morphology: word forms
- ▶ Syntax: word order
- ▶ Semantics: word sense, idioms

▶ 25

of 31

▶ 26

of 31

Required phrase-based SMT software is freely available

- ▶ Word alignment GIZA++
- ▶ Language modeling: SRILM, IRSTLM
- ▶ Phrase extraction: THOT, Moses
- ▶ Decoder: Moses

▶ 27

of 31

MT advances are published both in specific and NLP general conference and journals

- ▶ CONFERENCES: ACL, EAMT, AMTA, EMNLP...
- ▶ JOURNALS: Machine Translation, Computational Linguistics...

Relevant bibliography for this course

- ▶ <http://www.statmt.org/book/>
- ▶ *Philipp Koehn* (2010): Statistical Machine Translation. Publisher: Cambridge University Press. ISBN-10: 0521874157

- ▶ <http://www.cs.jhu.edu/~alopez/esslli2010.html>

*Our course
includes some
slides from
Adam Lopez!*



Let's comment the bibliography document

Evaluation Campaigns allow to compare different translation systems

- ▶ WMT (European Languages, started in 2005)
- ▶ IWSLT (Asian and European Languages, started in 2003)
- ▶ NIST (Chinese-English, Arabic-English, Urdu-English, started in 2001)

Next Session is about SMT theory

PART 1

- ▶ Statistical approach to Machine Translation
- ▶ Language model
- ▶ Translation model: word alignment, phrase extraction and probabilities

PART 2

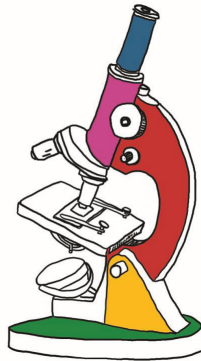
- ▶ Reordering
- ▶ Search
- ▶ Evaluation

The concept and feasibility of modern statistical machine translation

Statistical machine translation theory PART 1

Maxim Khalilov
TAUS Labs
Amsterdam

Marta R. Costa-jussà
Barcelona Media
Barcelona



RuSSIR 2012
August 5-10



Outline

- ▶ Statistical approach to Machine Translation
- ▶ Language model
- ▶ Translation model: word alignment, phrase extraction and probabilities

Statistical approach to Machine Translation



When I look at an article in Russian, I say: "this is really written in English, But it has been coded in some strange symbols. I will now proceed to decode."
Warren Weaver (1949)

▶ 3

of 28

Find most probable target sentence given a source language: $p(t | s)$

- ▶ Noisy channel

$$\hat{t} = \operatorname{argmax}_t p(t|s)$$

- ▶ Bayes rule

$$p(t|s) = \frac{p(t)p(s|t)}{p(s)}$$

- ▶ Final problem $\hat{t} = \operatorname{argmax}_t p(t)p(s|t)$

language model  translation model

What's a model?

For our purposes,

- ▶ A model will be a probability distribution over data
- ▶ Think of a probabilistic distribution as a story about how our data came into being
 - ▶ *We are learning so fast*
 - ▶ $p(\text{are}|\text{We})p(\text{learning}|\text{are})p(\text{so}|\text{learning})p(\text{fast}|\text{so})$

▶ 5

of 28

Divide and conquer

- ▶ Translation is faced by means of the joint optimization of:
 - ▶ **FLUENCY** : the language model focuses on keeping a fluent sentence while building the target sentence
 - ▶ **ADEQUACY** : the translation model focuses on keeping the meaning in the target sentence of the source sentence

▶ 6

of 28

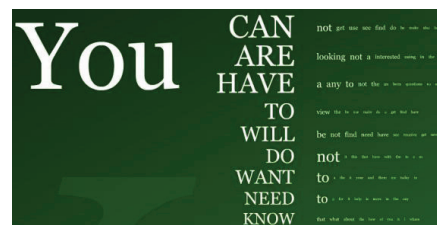
Outline

- ▶ Statistical approach to Machine Translation
- ▶ Language model
- ▶ Translation model: word alignment, phrase extraction and probabilities

▶ 7

of 28

Target language model



Unigram probabilities

$$p(w_1) = \frac{\text{count}(w_1)}{\text{total words observed}}$$

Bigram probabilities

$$p(w_2|w_1) = \frac{\text{count}(w_1w_2)}{\text{count}(w_1)}$$

Trigram probabilities

$$p(w_3|w_1w_2) = \frac{\text{count}(w_1w_2w_3)}{\text{count}(w_1w_2)}$$

▶ 8

of 28

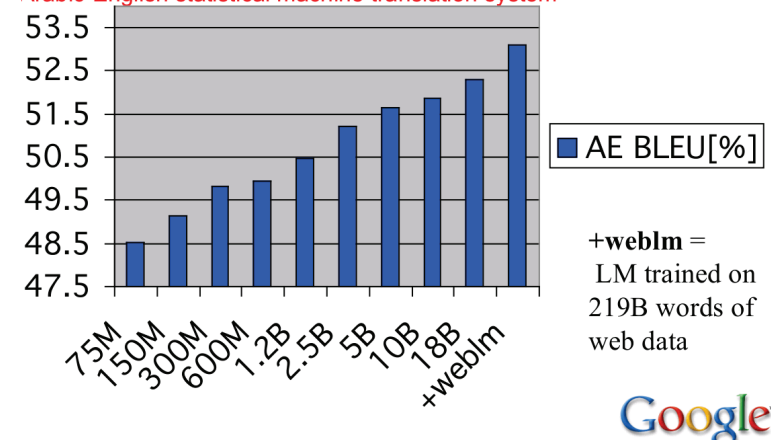
Sparse counts are a big problem

- ▶ Backing off avoids zero probabilities

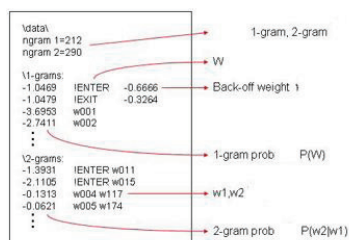
$$\begin{aligned}
 &.8 * p(w_3 | w_1 w_2) \\
 &+.15 * p(w_3 | w_2) \\
 &+.049 * p(w_3) \\
 &+.001
 \end{aligned}$$

How much data do we need to train a language model?

Impact on size of language model training data (in words) on quality of Arabic-English statistical machine translation system



Standard arpa format



Outline

- ▶ Statistical approach to Machine Translation
- ▶ Language model
- ▶ Translation model: word alignment, phrase extraction and probabilities

What is a Parallel Text or Parallel Corpus?

- ▶ Translated text/documents in two languages
- ▶ Ideally sentence-aligned

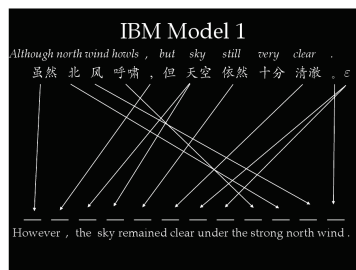
Table 2
Output from alignment program.

English	French
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates.	Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.
The higher turnover was largely due to an increase in the sales volume.	La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes.
Employment and investment levels also climbed.	L'emploi et les investissements ont également augmenté.

▶ 13

of 28

A good story for translation models...



IBM1

 $p(\text{English}|\text{Chinese})?$ $p(\text{English length}|\text{Chinese length})$ $p(\text{Chinese word position})$ $p(\text{English word}|\text{Chinese word})$

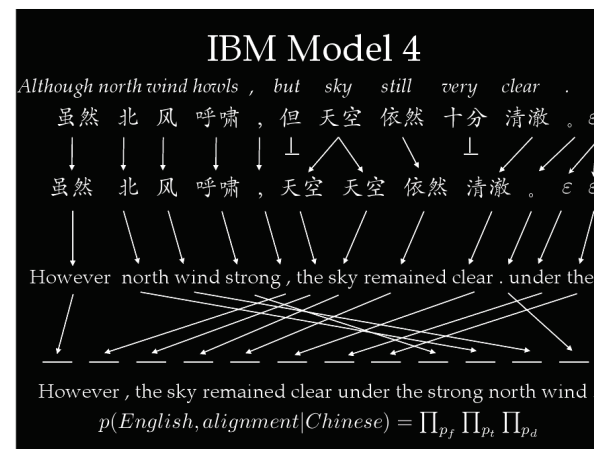
More complex models:

IBM2, IBM3, IBM4, IBM5

▶ 14

of 28

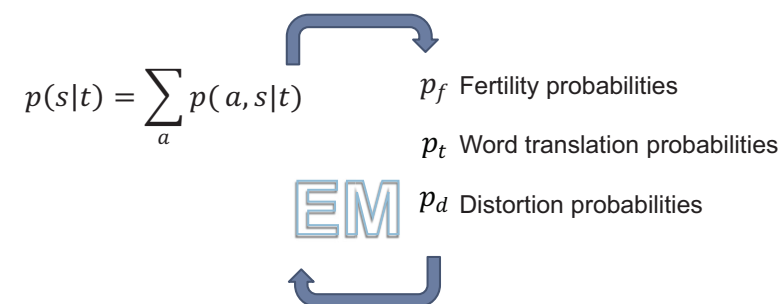
More about the translation story



▶ 15

of 28

Training the probabilities through EM



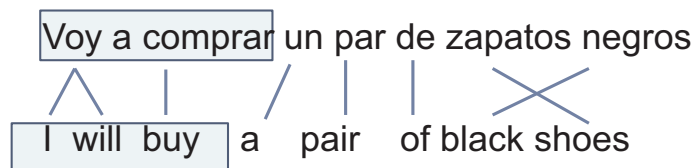
▶ Major problems:

- ▶ Weak reordering model...
- ▶ Many decisions at a time...

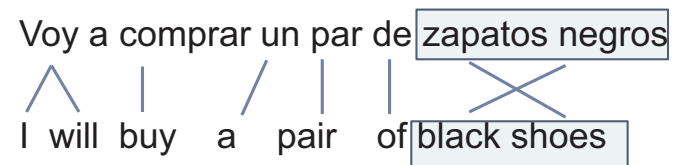
▶ 16

of 28

One step further: phrase-based models



One step further: phrase-based models



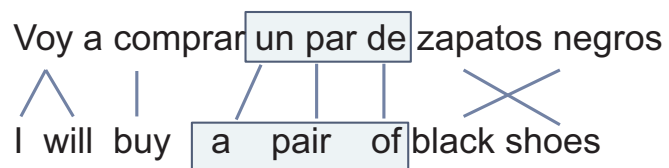
► 17

of 28

► 19

of 28

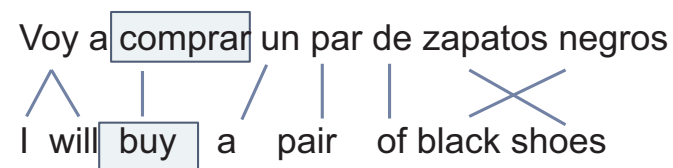
One step further: phrase-based models



**THESE ARE NOT PHRASES
IN ANY LINGUISTIC SENSE**

**Restriction to linguistic phrases seems
to hurt (Koehn et al. 2003)**

One step further: phrase-based models



Phrases of maximum length=1

a # NULL comprar # buy
a # un par # pair de # of
zapatos#shoes black # negros

NOT A PHRASE:

Voy # I

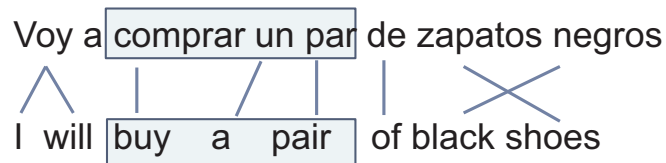
► 18

of 28

► 20

of 28

One step further: phrase-based models



Phrases of maximum length=3

Voy a comprar # I will buy a comprar un # buy a
comprar un par # buy a pair
de zapatos negros # of black shoes
...

NOT A PHRASE:

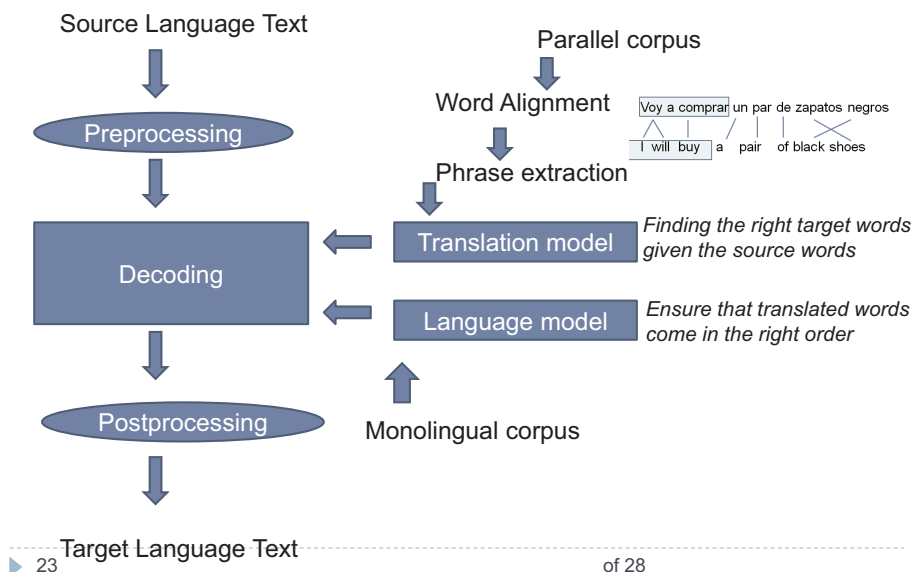
par de zapatos # pair of shoes

21

of 28

From previous session

A picture is worth a million equations



23

of 28

From previous session

A picture is worth a million equations

$$p(a, s|t) = \alpha(t) \prod_{j=1}^m p(s_j | t_i)$$

p(segmentation) → p(translation) p(reordering inside phrase)

$$p(s_j | t_i) = \frac{\text{count}(s_j, t_i)}{\text{count}(s_j)}$$

- Segmentation probabilities
- Phrase translation probabilities
- Distortion probabilities

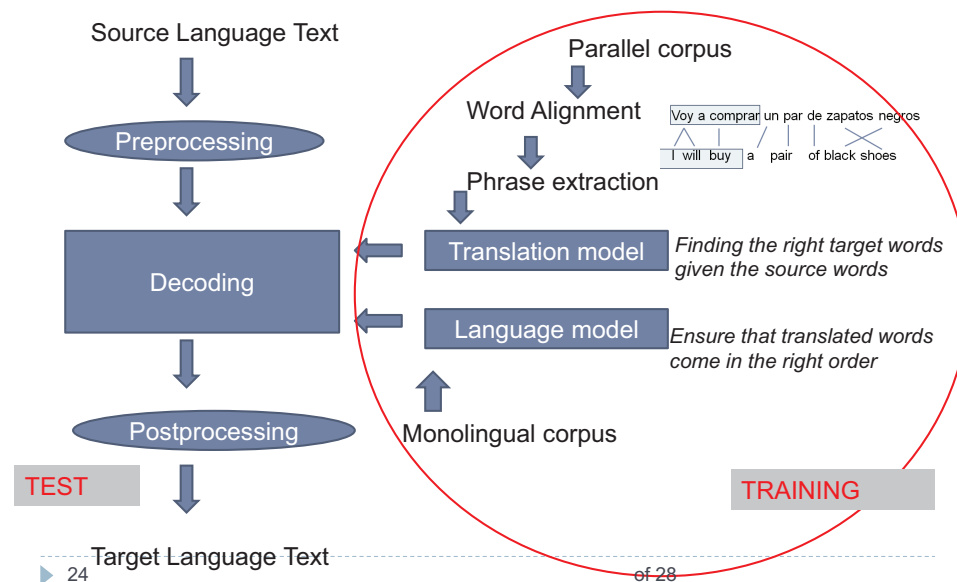
Major problems: weak reordering model

22

of 28

From previous session

A picture is worth a million equations



24

of 28

Log-linear combination of models: adding more features to noisy channel

- ▶ It constitutes a more general approach, which is based on maximum entropy principles (Berger et al., 1996)

$$\begin{aligned}\hat{t} &= \operatorname{argmax}_t \prod_i p_i(s, t)^{\lambda_i} \\ &= \operatorname{argmax}_t \sum_i \lambda_i \log(p_i(s, t))\end{aligned}$$

Note the noisy channel is a particular case:

$$\hat{t} = \operatorname{argmax}_t p(t)p(s|t)$$

Are you convinced with Phrase-based models?



- ▶ Phrase-based models are dumb:
 - ▶ No semantics, no syntax, no morphology...
- ▶ But... they are still regarded as state-of-the-art.
- ▶ Why? Simple models are easier to learn and deploy
- ▶ Need proof? Google uses a phrase-based model

▶ 25

of 28

Feature functions

- ▶ Lexical models, helpful when translation units are sparse
- ▶ Reordering model, used to provide reordering between phrases (**DETAILS ON THIS... LATER**)
- ▶ Word bonus, used to compensate the language model which benefits shorter outputs

Loglinear is clearly a misnomer as many features are not logarithms at all

▶ 26

of 28

▶ 27

of 28

The session continues... more SMT theory

- ▶ Statistical approach to Machine Translation
- ▶ Noisy channel:
 - ▶ language and translation model
 - ▶ word alignment, phrase extraction and probabilities
- ▶ Reordering
- ▶ Search
- ▶ Evaluation

**Just keep tuned !!!!!
It can get more
complicated...**



▶ 28

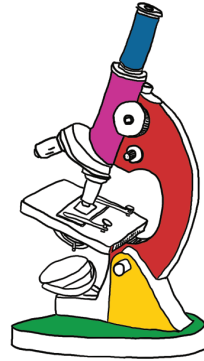
of 28

The concept and feasibility of modern statistical machine translation

Statistical machine translation theory **PART 2**

Maxim Khalilov
TAUS Labs
Amsterdam

Marta R. Costa-jussà
Barcelona Media
Barcelona



RuSSIR 2012
August 5-10, 2012

