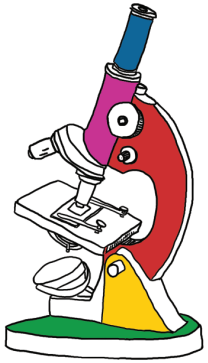
The concept and feasibility of modern statistical machine translation

Statistical machine translation theory
PART 2

Maxim Khalilov
TAUS Labs
Amsterdam

Marta R. Costa-jussà
Barcelona Media
Barcelona



RuSSIR 2012
August 5-10, 2012

Outline

- Decoding
- Reordering
- Evaluation of MT quality

Outline

- **Decoding**
- Reordering
- Evaluation of MT quality

Decoding

- Advances and the problem
- Decoding process
- Decoding: limiting reordering
- Decoding: errors

Decoding

- **Advances and the problem**
- Decoding process
- Decoding: limiting reordering
- Decoding: errors

Advances

- From "noisy channel" to log-linear combination *(Och and Ney, 2002)*

$$\hat{T} = \arg\max_T p(T \mid S) \approx \arg\max_T \prod_i P_i(S,T)^{\lambda_i}$$

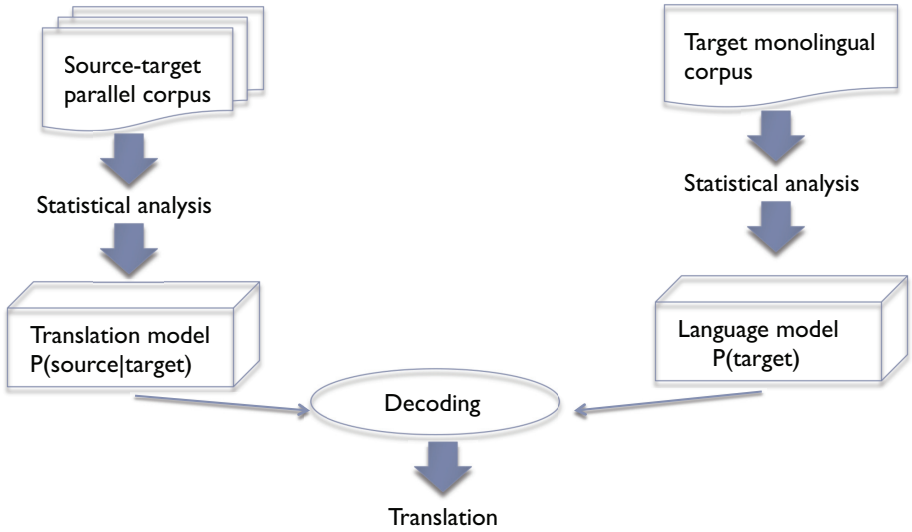
In this case, "noisy channel" is considered a particular case:

$$p_1(S,T) = p(S \mid T), p_1(S,T) = p(T), \lambda_1 = \lambda_2 = 1$$

- From word-based models to phrase-based models (Zens et al., 2002; Koehn et al., 2003)

$$p(S \mid T) = \frac{N(S,T)}{N(S)}$$

The problem



Decoding

- Advances and the problem
- **Decoding process**
- Decoding: limiting reordering
- Decoding: errors

Decoding: how to find the translation?

- Modeling problem: what is a good translation?
- Search problem: given a model and a source sentence how to find the translation that the model likes best?
- Task: to explore the space of possible translations using a search algorithm.

Decoding process

Spa: Maria no dio una botefada a la bruja verde

Eng: Maria did not slap the green witch

Decoding process

Build translation from left to right

Decoding process

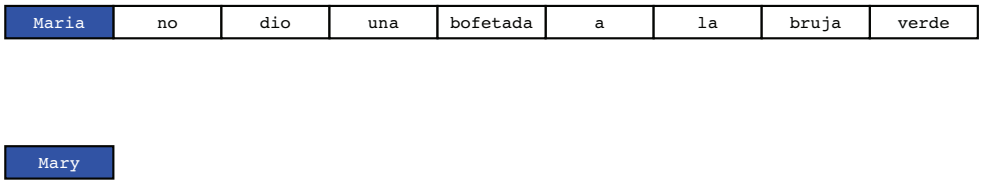
Build translation from left to right:

1. Select source words to be translated
2. Find target phrase translation
3. Add target phrase to end of partial translation

Decoding process

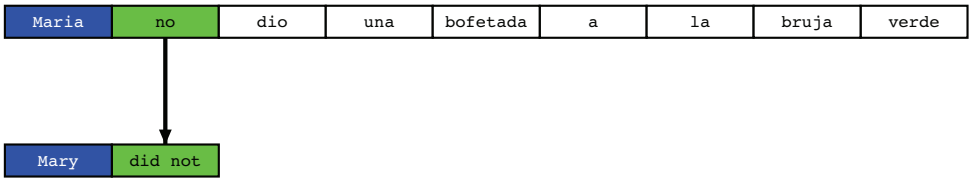
Build translation from left to right:

1. Select source words to be translated
2. Find target phrase translation
3. Add target phrase to end of partial translation
4. Mark source words as translated



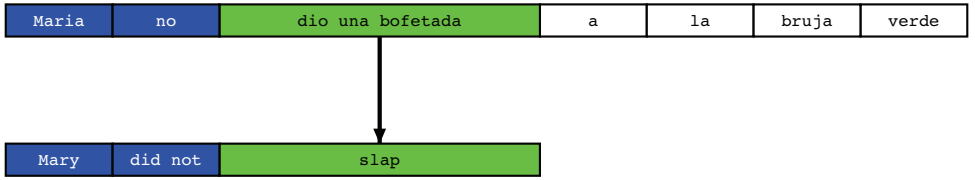
Decoding process

- One to many translation



Decoding process

- Many to one translation



Decoding process

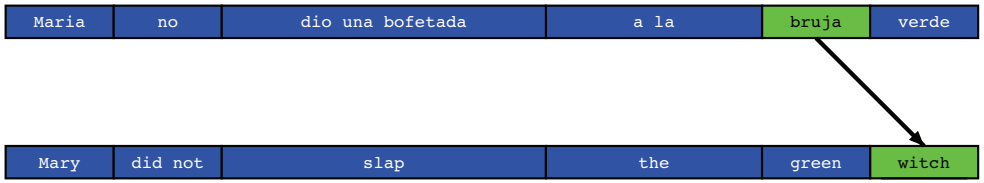
- Many to one translation

Decoding process

- Reordering

Decoding process

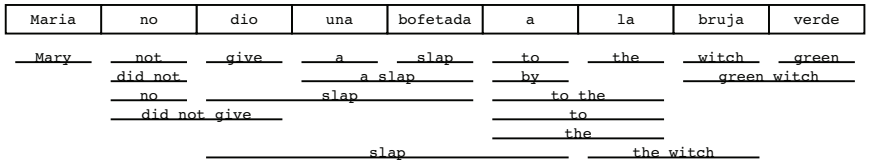
- Reordering



- And translation finished

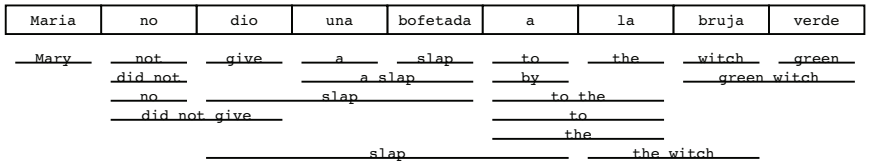
Decoding: hypothesis expansion

Let's look up possible phrase translation options:



- Many different ways to segment words into phrases
- Many different ways to translate each phrase
- Many different ways to reorder phrases

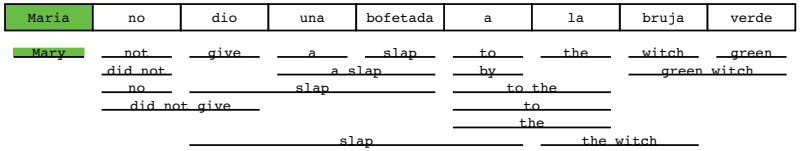
Decoding: hypothesis expansion



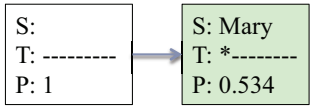
Start with empty hypothesis:
- no source words covered
- no target words covered
- probability is 1

S:
T: ---------
P: 1

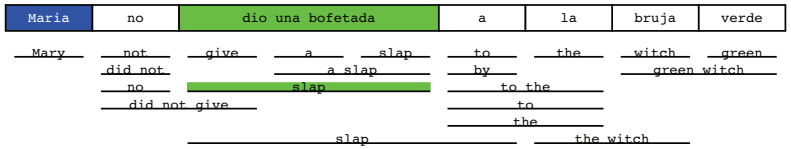
Decoding: hypothesis expansion



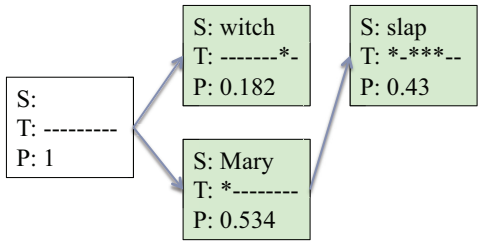
Pick translation option
Create translation hypothesis:
- S: add source phrase Mary
- T: first target word covered
- P: probability 0.534



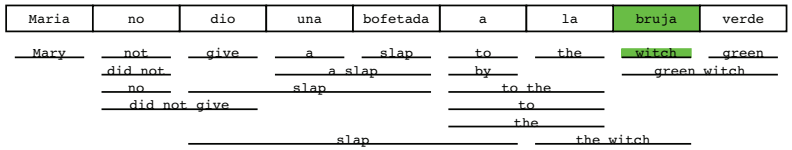
Decoding: hypothesis expansion



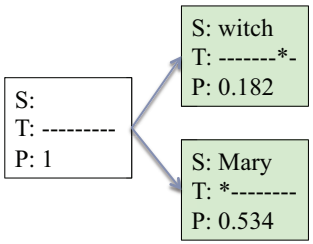
Further hypothesis expansion



Decoding: hypothesis expansion

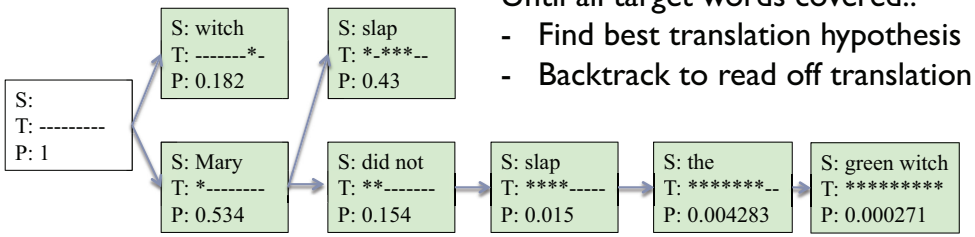


Add another hypothesis

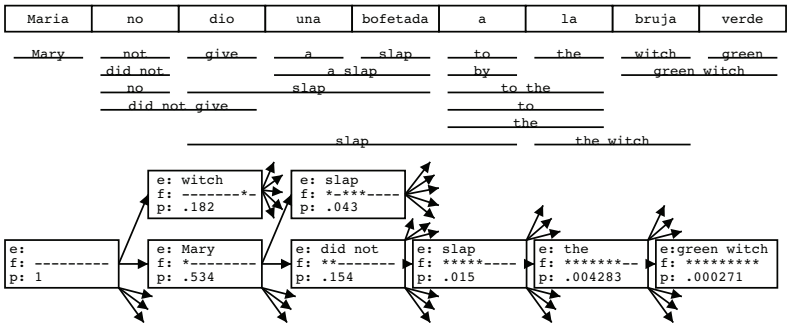


Decoding: hypothesis expansion

Until all target words covered..
- Find best translation hypothesis
- Backtrack to read off translation



Decoding: hypothesis expansion



Explosion of search space

Decoding: complexity

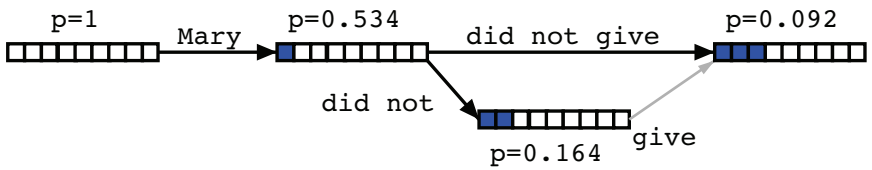
- This is a BIG search problem:
  - - segmentation – $O(2^n)$
  - - substitutions - $O(5^n)$
  - - permutations – $O(n!)$
- Possible solutions:
  - - Dynamic programming
  - - Approximation (beam search)
  - - Model restrictions (reordering)

Decoding: complexity

- Can we do it more efficiently than $O(5n^2 2^n)$??
- No!! Knight (1999) shows that this task is NP-Complete
- There is need to reduce search space
  - - risk free strategy: hypothesis recombination
  - - risky strategy: histogram/threshold pruning

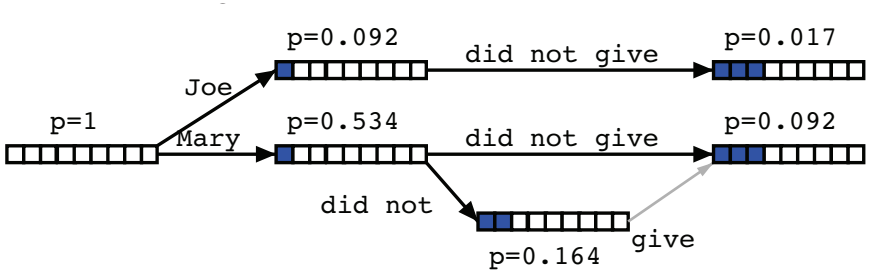
Decoding: hypothesis recombination

- Different paths to the same partial hypothesis:
  - Combine paths: drop weaker paths
  - Keep pointer from the weaker paths



Decoding: hypothesis recombination

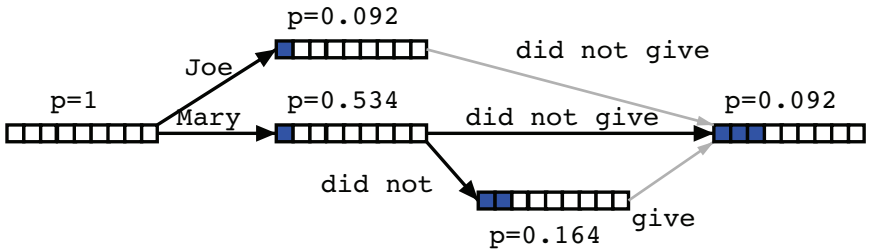
- Recombined hypotheses do not have to match completely
- No matter what is added, weaker paths can be dropped, if
  - - last two target words match (for LM)
  - - source word coverage vectors match (affect future path)



Decoding: hypothesis recombination

- Recombined hypotheses do not have to match completely
- No matter what is added, weaker paths can be dropped, if
  - - last two target words match (for LM)
  - - source word coverage vectors match (affect future path)

  => Combine paths!



Decoding: pruning

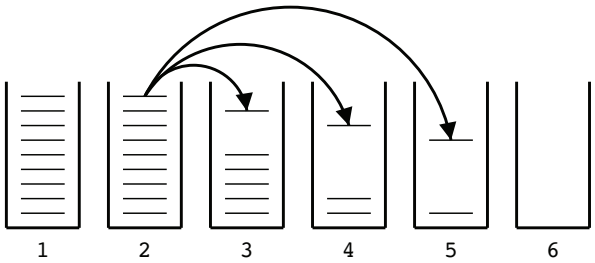
- But! Hypothesis recombination is not sufficient. Possible solution is to further reduce search space with approximation (pruning). We can heuristically discard weak hypotheses.
- Idea: to prune states by accumulated path length
- Organize hypotheses in stacks by:
  - - same source words covered
  - - same number of source words covered
  - - same number of target words generated

Decoding: pruning

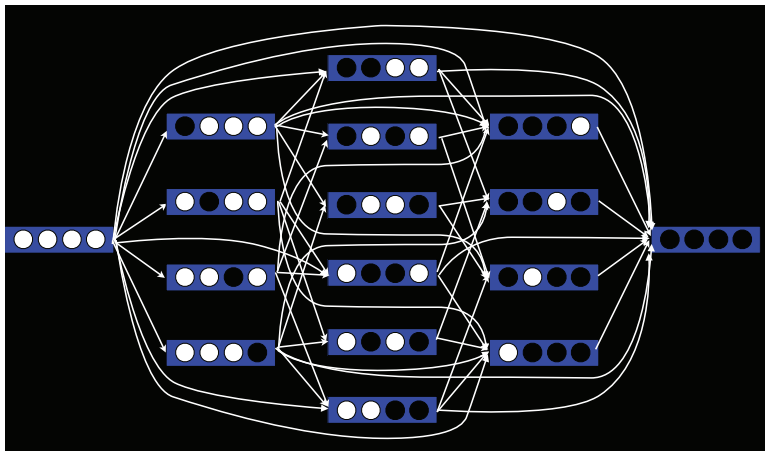
- Compare hypotheses in stacks, discard bad ones
  - Histogram pruning: keep top *n* hypotheses in each stack (e.g., *n=1000*)
  - Threshold pruning: keep hypotheses that are at most *k* times the cost of best hypothesis in stack (e.g., *k=0.001*)

Decoding: pruning

- "Stack decoding": a linear-time approximation
- Organization hypotheses into stacks:
  - - based on number of source words translated
  - - during translation all hypotheses from one stack are expanded
  - - expanded hypotheses are placed into stacks

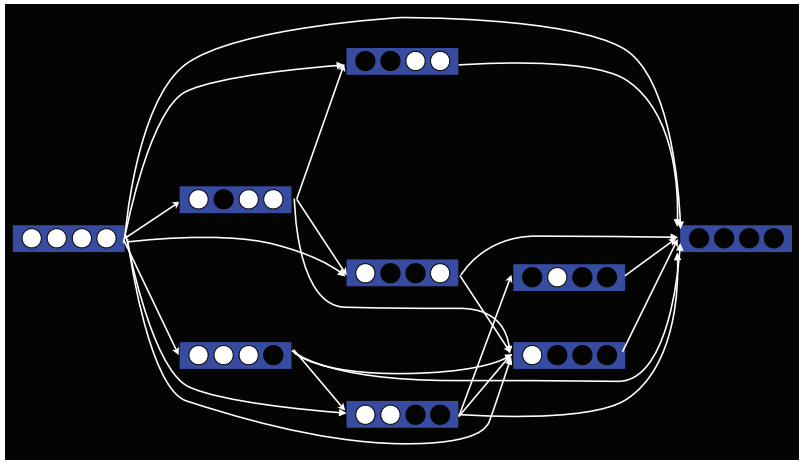


Decoding: pruning



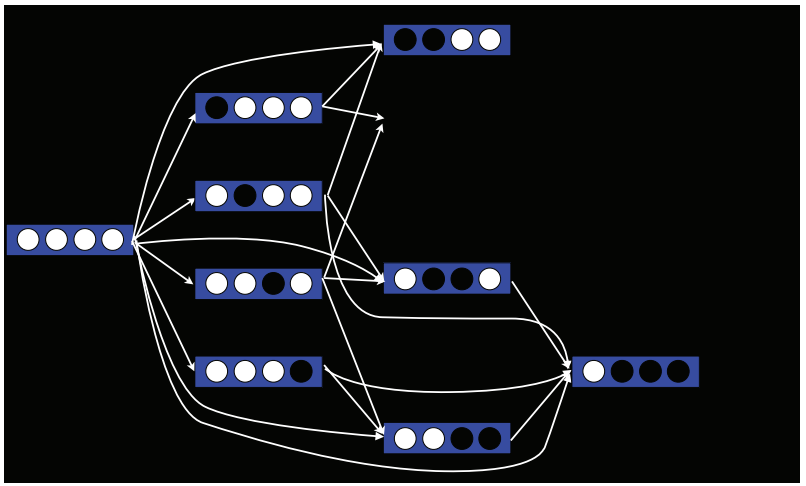
Decoding: pruning

- Prune states by accumulated path length



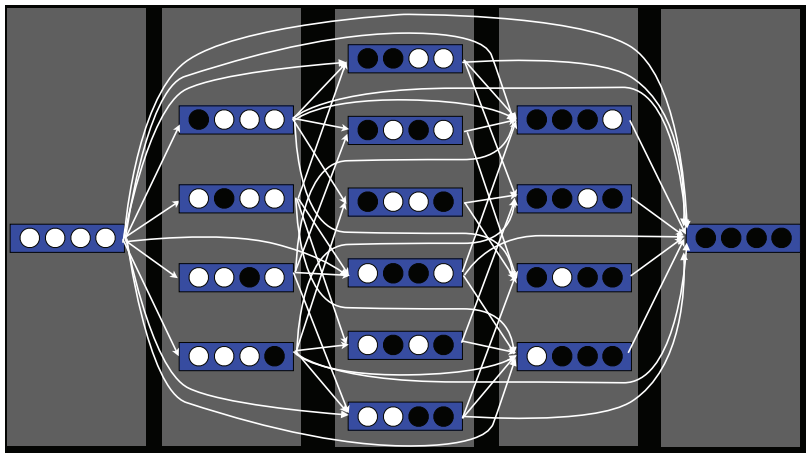
Decoding: pruning

- Reality: longer paths have lower probability



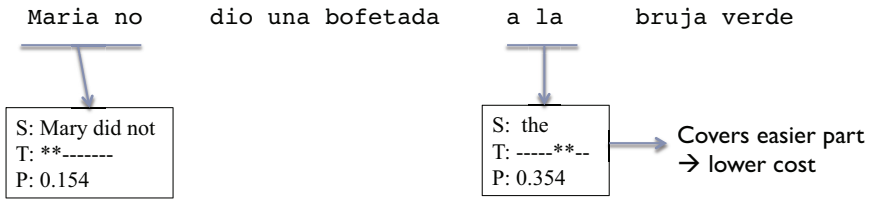
Decoding: pruning

- Solution: group states by number of covered words



Decoding: pruning

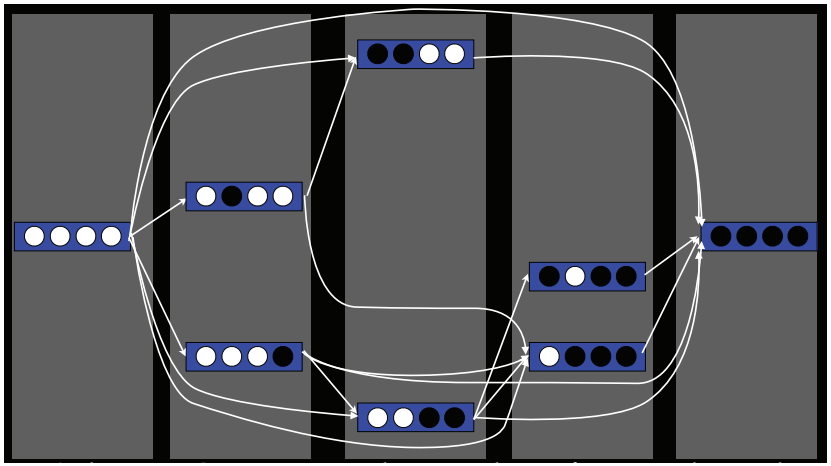
- How to compare hypotheses with same number of source words covered?



- Hypotheses that cover easy part of the sentence are preferred
- Need to consider a future cost of uncovered parts to estimate cost to translate the remaining part of current input

Decoding: pruning

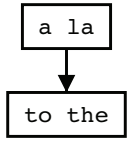
- "Stack decoding": a linear-time approximation



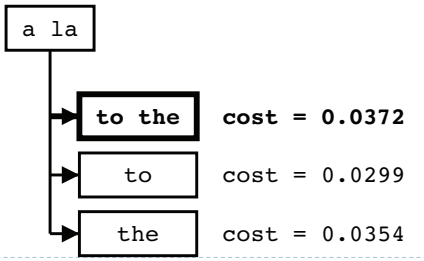
Decoding: pruning

- Future cost estimation:
- Step 1: estimate future cost for each translation option
  - Look up TM cost
  - Estimate LM cost (no prior context)
  - Ignore reordering cost

    ->LM*TM=p(to)*p(the|to)*p(to the|a la)



- Step 2: find cheapest cost among translation options

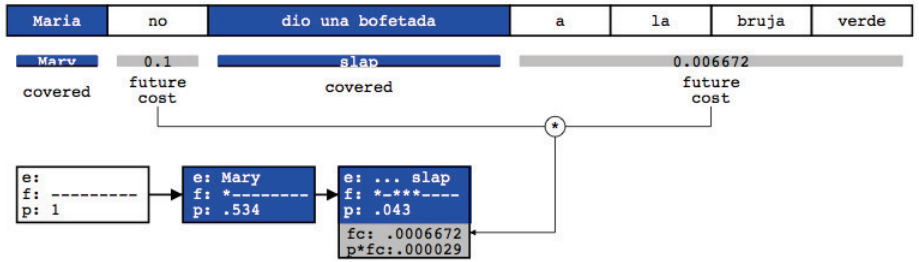


Decoding: pruning

- Step 3: find cheapest future cost for each possible span.
- Cheaper cost can be either:
  - Cost of translation option for that span, or
  - Sum of costs of covering subspans
- Pre-computing future costs, bottom up: a dynamic programming solution.

Decoding: pruning

- Application of future costs when pruning hypotheses
- For each uncovered continuous span:
  - Look up future costs for each maximal contiguous uncovered span
  - Add to actually accumulated cost for translation option for pruning



Decoding

- Advances and the problem
- Decoding process
- **Decoding: limiting reordering**
- Decoding: errors

Decoding: limiting reordering

- Reordering may be limited:
  - Monotone translation: no reordering at all
  - Only phrase movements of at most *n* words
- Reordering limits speed up search (polynomial instead of exponential)

More about reordering – after this session

Decoding

- Advances and the problem
- Decoding process
  - Hypothesis expansion
  - Complexity
  - Hyptothesis recombination
  - Pruning
- Decoding: limiting reordering
- **Decoding: errors**

Decoding: errors

- Search errors: there was a higher scoring translation, but we failed to find it.
- Model errors: the models assigns lower probability to better translation.

Decoding: n-best list

```
Translation ||| Reordering LM TM WordPenalty ||| Score
this is a small house ||| 0 -27.0908 -1.83258 -5 ||| -28.9234
this is a little house ||| 0 -28.1791 -1.83258 -5 ||| -30.0117
it is a small house ||| 0 -27.108 -3.21888 -5 ||| -30.3268
it is a little house ||| 0 -28.1963 -3.21888 -5 ||| -31.4152
this is an small house ||| 0 -31.7294 -1.83258 -5 ||| -33.562
it is an small house ||| 0 -32.3094 -3.21888 -5 ||| -35.5283
this is an little house ||| 0 -33.7639 -1.83258 -5 ||| -35.5965
this is a house small ||| -3 -31.4851 -1.83258 -5 ||| -36.3176
this is a house little ||| -3 -31.5689 -1.83258 -5 ||| -36.4015
it is an little house ||| 0 -34.3439 -3.21888 -5 ||| -37.5628
it is a house small ||| -3 -31.5022 -3.21888 -5 ||| -37.7211
this is an house small ||| -3 -32.8999 -1.83258 -5 ||| -37.7325
it is a house little ||| -3 -31.586 -3.21888 -5 ||| -37.8049
this is an house little ||| -3 -32.9837 -1.83258 -5 ||| -37.8163
the house is a little ||| -7 -28.5107 -2.52573 -5 ||| -38.0364
the is a small house ||| 0 -35.6899 -2.52573 -5 ||| -38.2156
is it a little house ||| -4 -30.3603 -3.91202 -5 ||| -38.2723
the house is a small ||| -7 -28.7683 -2.52573 -5 ||| -38.294
it 's a small house ||| 0 -34.8557 -3.91202 -5 ||| -38.7677
this house is a little ||| -7 -28.0443 -3.91202 -5 ||| -38.9563
it 's a little house ||| 0 -35.1446 -3.91202 -5 ||| -39.0566
this house is a small ||| -7 -28.3018 -3.91202 -5 ||| -39.2139
```


Outline

- Decoding
- **Reordering**
- Evaluation of MT quality

Reordering

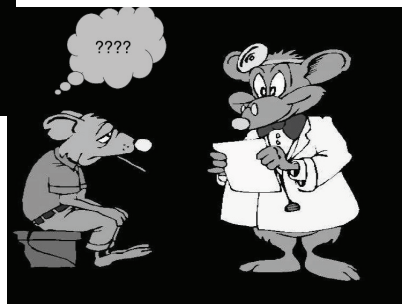
- Motivation and classification
- Constrained distance-based reordering
- Reordering as pre-processing
- Why syntax can be useful?
- Other significant works

Reordering

- **Motivation and classification**
- Constrained distance-based reordering
- Reordering as pre-processing
- Why syntax can be useful?
- Other significant works

Word reordering: motivation

Machine translation:
"Los resultados muestran un cierto diagnóstico"



Wrong order -> wrong translation:
"*Unspecified diagnosis*"

Correct translation:
"*true diagnosis*" ("diagnóstico cierto")

Word reordering: motivation

Original sentence:
"De behandeling kan niet genoeg beklemtoond worden"

Machine translation (without reordering model):
"The treatment cannot enough empasized be"

Correct translation:
"The treatment cannot be empasized enough"

Wrong order -> "word salad"

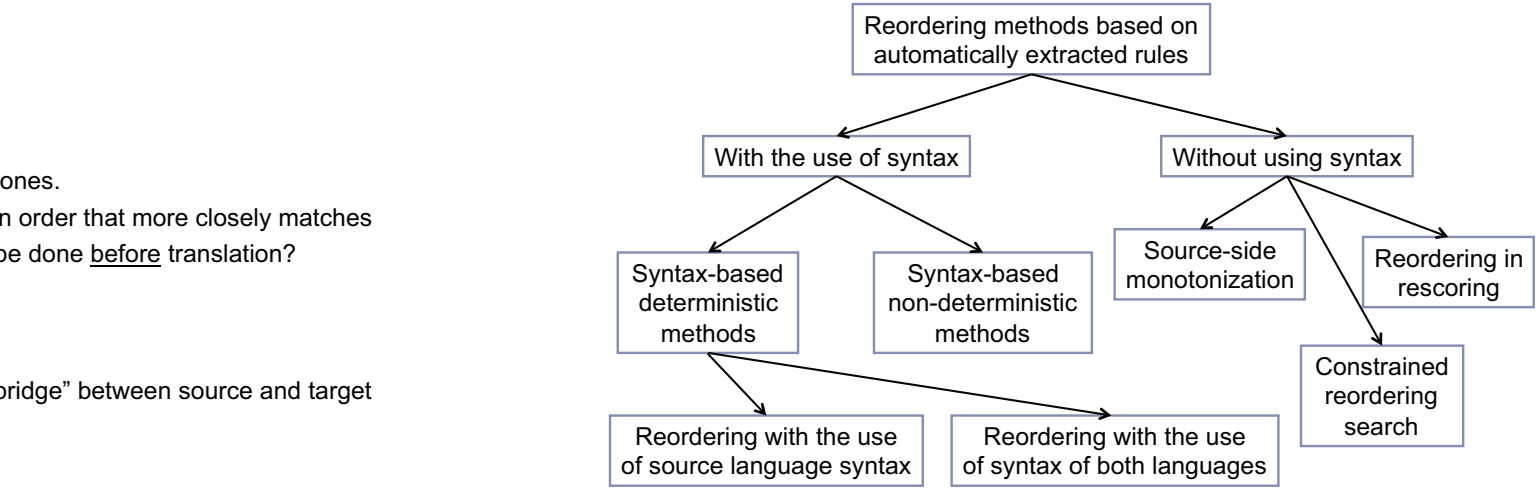
Word reordering and SMT process

- SMT process:

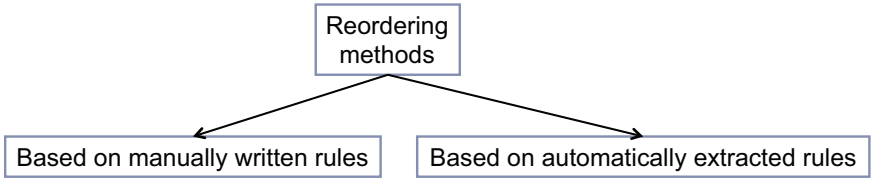
  1. Segment source-side input.
  2. Translate source words into target ones.
  3. Place translated source words in an order that more closely matches that of the target language. Can it be done before translation?

  Word alignment is usually used as a “bridge” between source and target languages

State-of-the-art reordering methods



State-of-the-art reordering methods



Reordering models based on manually written rules:

- (Collins et al., 2005) – a German parse tree is used for moving German verbs towards the beginning of the clause.
- (Popovic and Ney, 2006) - POS tag information is used to rewrite the input sentence between Spanish-English and German-English language pairs.
- (Zwarts and Dras, 2007) the natural language tendency to minimize the distance between a head and its dependents derived from the dependency trees is exploited to automatically reorder source-side constituents.

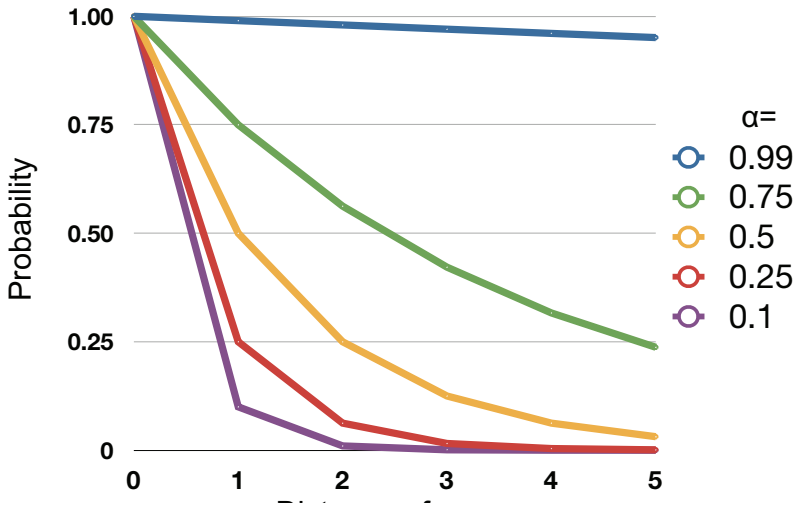
State-of-the-art reordering methods

- Non-deterministic approach:
  +: decoder has access to multiple reordering options
  -: the reordering search space can be huge
- Deterministic approach:
  +: simplicity and compatibility
  -: hard decision about word order (reordering mistakes cannot be corrected during decoding)
- A two-step integrated approach
  General idea: first, to permute the source words to account for global phenomena as local, second, to attack the local reordering problem with an established non-deterministic technique

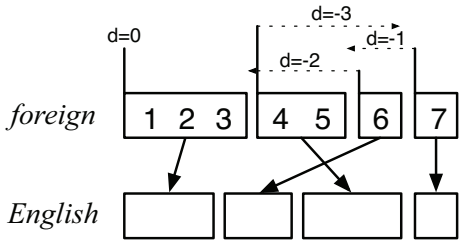
Reordering

- Motivation and classification
- **Constrained distance-based reordering**
- Reordering as pre-processing. Why syntax can be useful?
- Other significant works

Distance-based reordering



Distance-based reordering



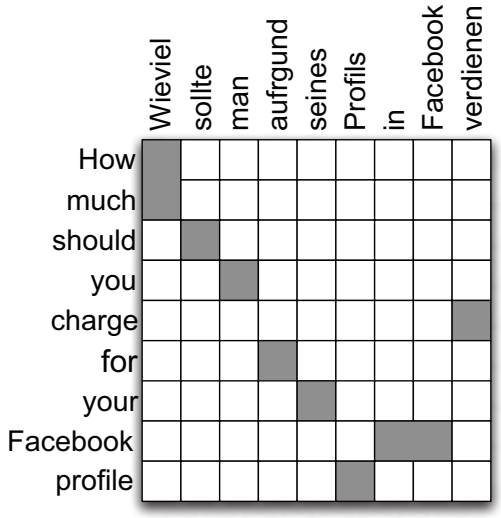
phrase	translates	movement	distance
1	1–3	start at beginning	0
2	6	skip over 4–5	+2
3	4–5	move back over 4–6	-3
4	7	skip over 6	+1

Scoring function: $d(x) = \alpha^{|x|}$ – exponential with distance

Distance-based reordering

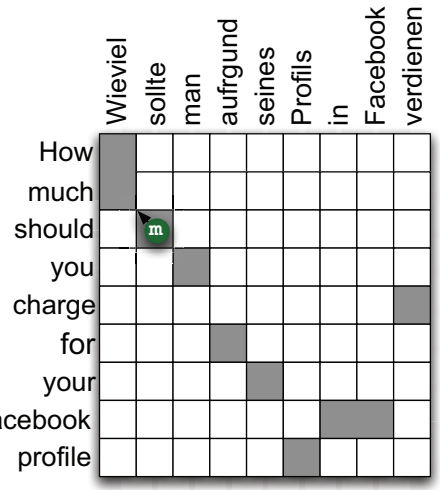
- Small values of α, severely discourage reordering
  - –Limit reordering to monotonic or a narrow window
  - –OK for languages with very similar word orders
  - –Bad for languages with different word orders
- The distance-based penalty applies uniformly to all words and all word types
  - –Doesn't know that adjectives and nouns should swap when translating from French to English
- Puts most responsibility on the language model

Lexicalized reordering (MSD)



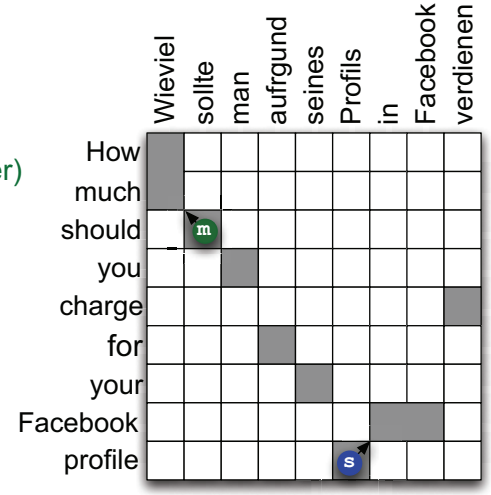
Lexicalized reordering (MSD)

m: monotone (keep order)



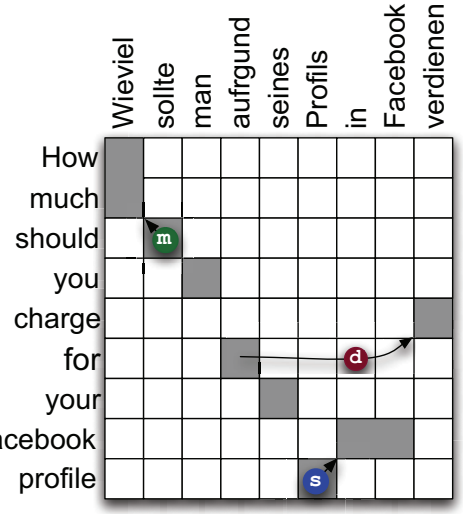
Lexicalized reordering (MSD)

m: monotone (keep order)
s: swap order



Lexicalized reordering (MSD)

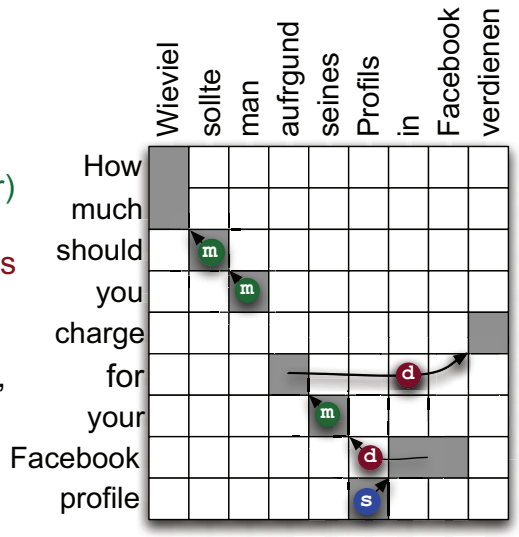
m: monotone (keep order)
s: swap order
d: become discontinuous



Lexicalized reordering (MSD)

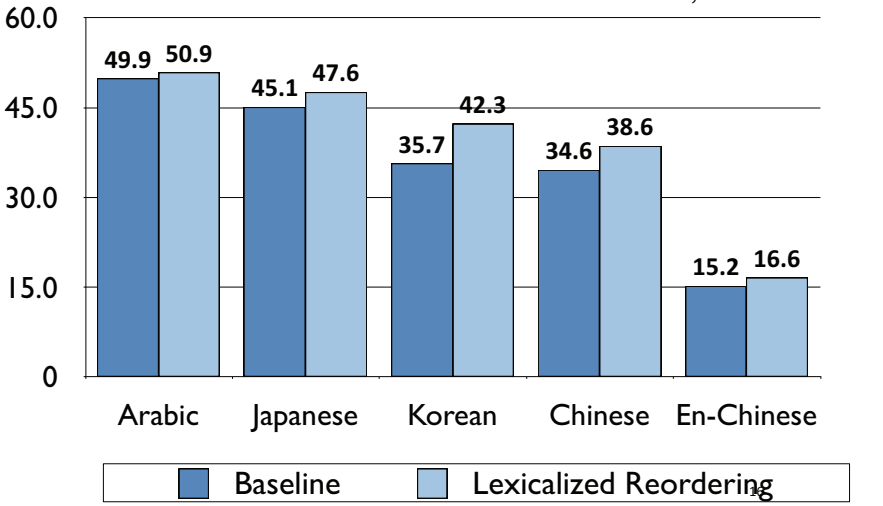
m: monotone (keep order)
s: swap order
d: become discontinuous

Reordering features are probability estimates of s, d, and m



Lexicalized reordering (MSD)

Koehn et al, IWSLT 2005



Lexicalized reordering (MSD)

- Identical phrase pairs <f,e> as in the phrase translation table
- Contains values for p(monotone|e,f), p(swap|e,f), p(discontinuous|e,f)

Source	Translation	p(m|e,f)	p(s|e,f)	p(d|e,f)
natuerlich	of course	0.52	0.08	0.40
natuerlich	naturally	0.42	0.10	0.48
natuerlich	of course ,	0.50	0.001	0.499
natuerlich	, of course	0.27	0.17	0.56

Constrained reordering search

Reordering constraints aim to limit the search space with minimal loss of generality during decoding:

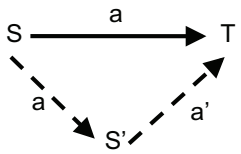
- **IBM** constraint - make the search feasible by introducing restrictions of the search space at the word level in the spirit of the IBM constraints
- **ITG** (Inverse Transduction Grammar) constraint: the input sentence is interpreted as a sequence of word blocks. For each two adjacent blocks, a decision is taken either to invert the original order, or to leave it as is.
- **A maximum entropy** model, transforming the reordering prediction into a classification problem.
- The **local** constraint: a simplification of the IBM constraint allowing for local permutations only.
- The **MaxJumps** constraint numerically limits the number of reorderings specified by two parameters:
  - m - a maximum distance measured in words, that a source word can be reordered (a distortion limit)
  - j - a maximum number of "jumps" within a sentence (a reordering limit)

Reordering

- Motivation and classification
- Constrained distance-based reordering
- **Reordering as pre-processing. Why syntax can be useful?**
- Other significant works

Reordering as pre-processing

Idea: Why not to reorder before translation?

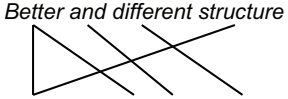


- Source reordering of *S* is as successful as much as the alignment *a'* is monotone.
- This problem can be seen as the task of learning from a word-aligned parallel corpus a model of source permutation from *S* to *S'*, where the latter has (almost) monotone alignment with *T*.

Deterministic method without use of syntax

Statistical machine reordering (SMR) approach proposed in (Costa-jussà, 2006):

Better and different structure



Estructura mejor y differente

Alignment: 1-1 1-2 2-3 3-4 4-1

1-2 2-3 3-4 4-1

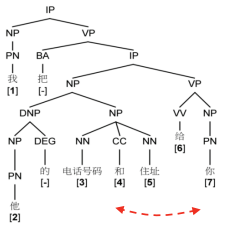
Reordered positions: 2 3 4 1

Classes: C36 C88 C23 C63

Why syntax can be useful?

Zh:	我	把	他	的	电话号码	和	住址	给	你
Gloss:	I	BA	his		telephone	and	address	give	you

Ref: *I give you his telephone number and address*

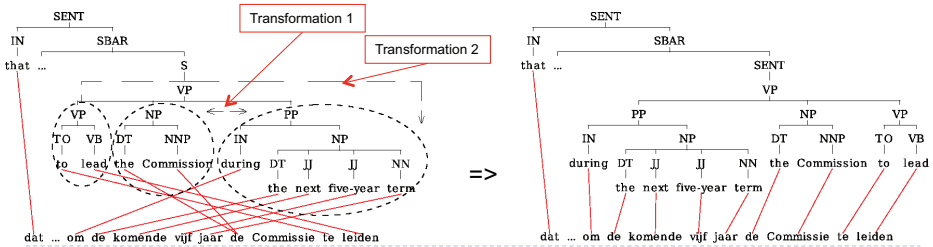


Why syntax can be useful?

Major challenges of English-to-Dutch translation:

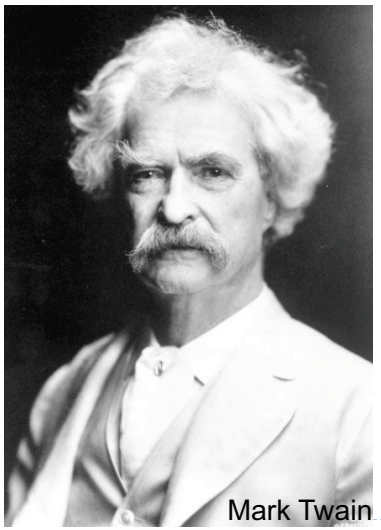
- Dutch verb appears in the end of the relative clause
- Phrasal verbs are different
- Differences in the positioning of adverbial structures

Source-side parse tree transformation:



German-English reordering (Collins)

“The Germans have another kind of parenthesis, which they make by splitting a verb in two and putting half of it at the beginning of an exciting chapter and the OTHER HALF at the end of it. Can any one conceive of anything more confusing than that? These things are called ‘separable verbs.’ The wider the two portions of one of them are spread apart, the better the author of the crime is pleased with his performance.”


Mark Twain

German-English reordering (Collins)

Ich	werde	Ihnen	den	Report	aushaendigen	.
I	will	to_you	the	report	pass_on	.

Ich	werde	Ihnen	die	entsprechenden	Anmerkungen	aushaendigen	.
I	will	to_you	the	corresponding	comments	pass_on	.

Ich	werde	Ihnen	die	entsprechenden	Anmerkungen	am	Dienstag	aushaendigen
I	will	to_you	the	corresponding	comments	on	Tuesday	pass_on

German-English reordering (Collins)

Main clause

Ich	werde	Ihnen	den	Report	aushaendigen	,
I	will	to_you	the	report	pass_on	,

Subordinate clause

damit	Sie	den	eventuell	uebernehmen	koennen	.
so_that	you	it	perhaps	adopt	can	.

German-English reordering (Collins)

Phrase-based models have an overly simplistic way of handling different word orders.

We can describe the linguistic differences between different languages.

Collins defines a set of 6 simple, linguistically motivated rules, and demonstrates that they result in significant translation improvements.

German-English reordering (Collins)

Step 1: Reorder the source language

Ich werde Ihnen den Report aushaendigen ,
damit Sie den eventuell uebernehmen koennen .

Ich werde aushaendigen Ihnen den Report ,
damit Sie koennen uebernehmen den eventuell .

(I will pass_on to_you the report, so_that you can adopt it perhaps .)

Step 2: Apply the phrase-based machine translation pipeline to the reordered input.

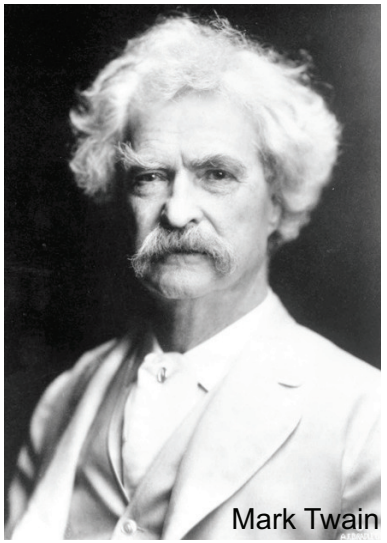
German-English reordering (Collins)

Ich werde Ihnen den Report aushaendigen, damit Sie den eventuell uebernehmen koennen.

Ich werde aushaendigen Ihnen den Report, damit Sie koennen uebernehmen den eventuell.

I will to_you the report pass_on, so_that you it perhaps adopt can.

I will pass_on to_you the report, so_that you can adopt it perhaps .



German-English reordering (Collins)

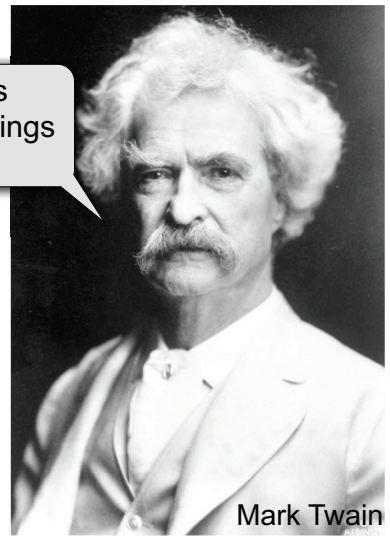
Ich werde Ihnen den Report aushaendigen, dam
eventuell uebernehmen koennen.

Ich werde aushaendigen Ihnen den Report, damit Sie koennen uebernehmen den eventuell.

I will to_you the report pass_on, so_that you it perhaps adopt can.

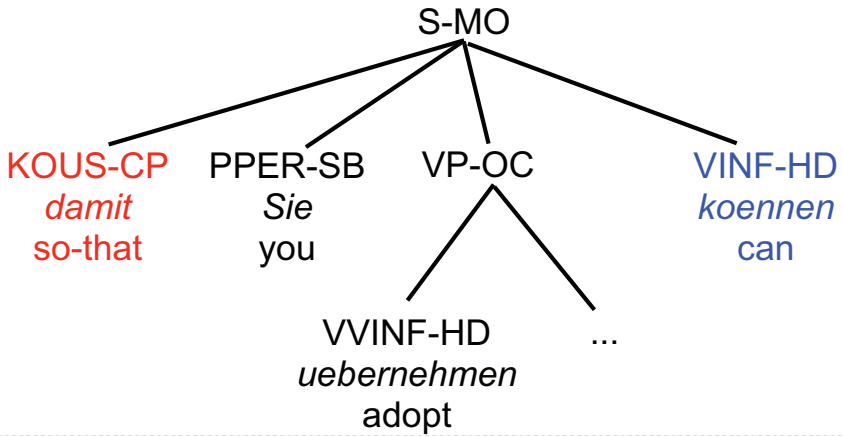
I will pass_on to_you the report, so_that you can adopt it perhaps .

Now that seems less like the ravings of a madman.

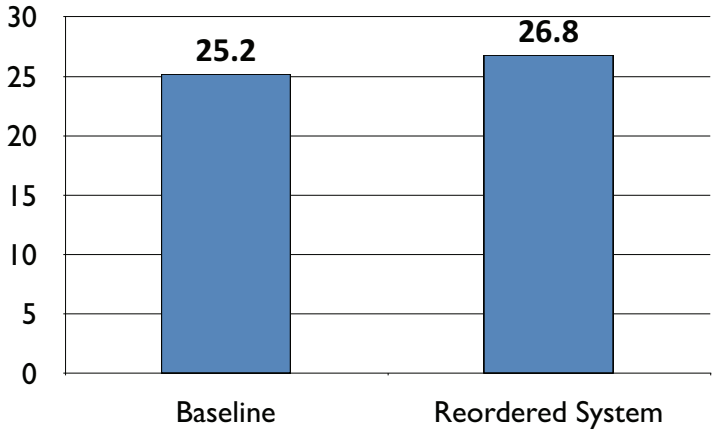


German-English reordering (Collins)

One of the rules: in a subordinate clause move the head of the clause to follow the complementizer



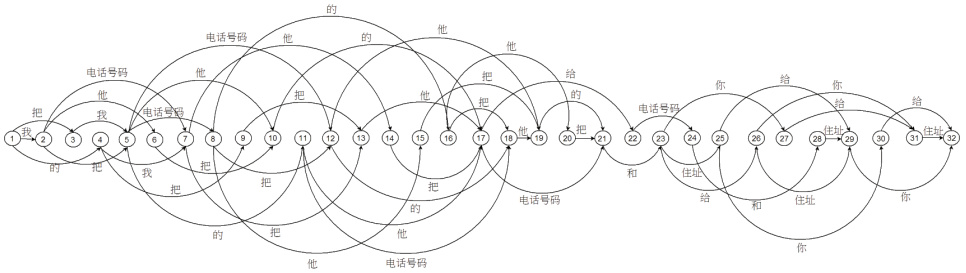
German-English reordering (Collins)



Significant improvement at p<0.01 using the sign test

Non-deterministic POS-based method

Proposed in (Crego, 2008):

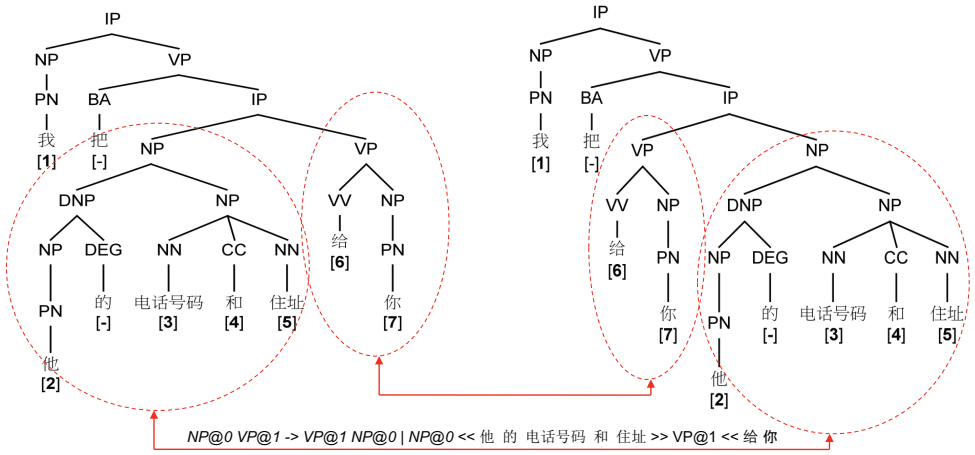


The number of possible permutations of the Chinese words is limited to the POS permutations seen in the training corpus.

S->S'->nxS'->T

n×S' is a word lattice, compactly representing the n-best reorderings of the source-side sentences S'.

Syntax-based reordering



我 把 他 的 电话号码 和 住址 给 你

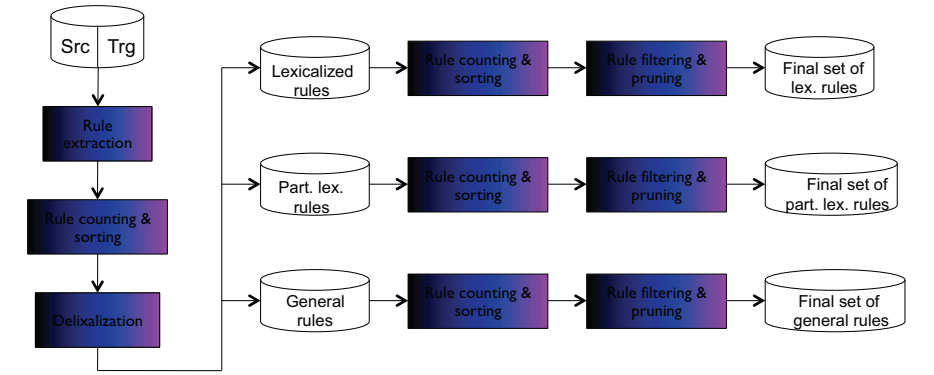
I BA his telephone and address give you

我 把 给 你 他 的 电话号码 和 住址

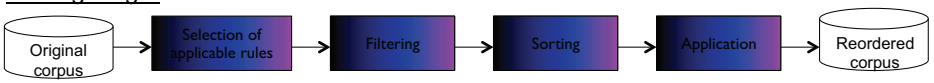
I BA give you his telephone and address

Syntax-based reordering

Training stage:



Testing stage:



Reordering

- Motivation and classification
- Constrained distance-based reordering
- Reordering as pre-processing
- Why syntax can be useful?
- **Other significant works**

Other significant works

- (Xia and McCord, 2004): a system for French-to-English translation based on the principle of automatic rewrite pattern extraction using an isomorphic parse tree and phrase alignments
- (Elming, 2008): syntactically motivated rewrite patterns are first combined in the weighted lattice of alternative translations, and then integrated in phrase-based SMT
- Modification of distortion matrix using chunk information (Bisazza and Federico, 2012)
- (Galley and Manning, 2006): an extension of the MSD model (Tillman, 2004) able to handle long-distance reorderings
- (Tromble and Eisner, 2009): reordering is seen as a learning problem of word permutations. The space of permutations is structured using a binary SCFG.
- (Yahyaei and Monz, 2010): improve the reordering space definition with a classifier guessing the most likely jump position

Outline

- Decoding
- Reordering
- **Evaluation of MT quality**

Evaluation of MT quality

- Human and automatic MT evaluation: major issues
- Automatic metrics
  - BLEU
  - METEOR, TER and GTM
- Human evaluation
- Remaining gaps

Evaluation of MT quality

- **Human and automatic MT evaluation: major issues**
- Automatic metrics
  - BLEU
  - METEOR, TER and GTM
- Human evaluation
- Remaining gaps

MT evaluation: major issues

- Evaluating MT output is not quite the same as evaluating human translation
  - Very different profile and characteristics of errors
  - Often MT is targeted for different purpose or use than human translation: different measures are required
- MT Evaluation is difficult:
  - Language variability – there is no single correct translation
  - Human evaluation is subjective
  - How good is "good enough"? Depends on task or application
  - Is system A better than system B? Depends on specific criteria…
- Some well-established methods, but no standard or single approach that is universally accepted
- MT Evaluation is still a research topic in itself!
  - How do we assess whether an evaluation method is good?

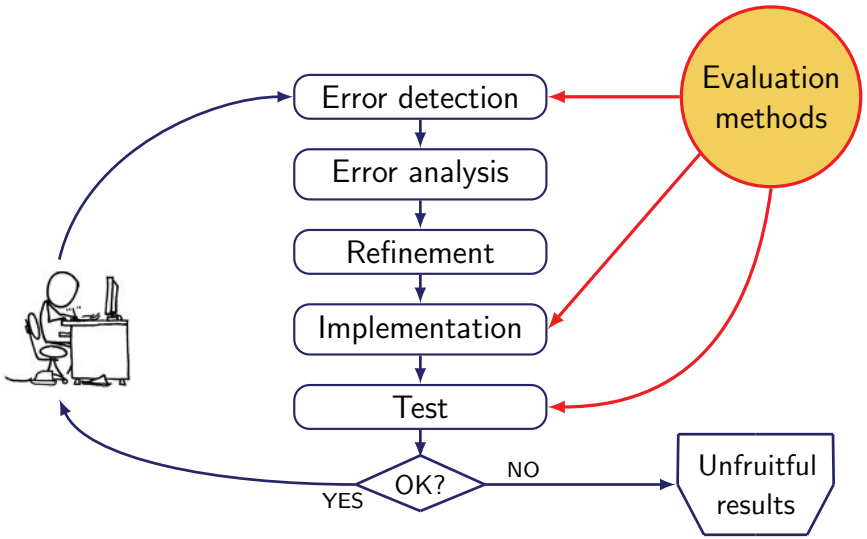
Dimensions of MT evaluation

- Human evaluation vs. automated metrics
- Quality assessment at sentence (segment) level vs. task-based evaluation
- "Black-box" vs. "Glass-box" evaluation
- Evaluation for external validation vs. target function for automatic system tuning vs. ongoing quality assessment of MT output

Evaluation of MT quality

- Human and automatic MT evaluation: major issues
- **Automatic metrics**
  - BLEU
  - METEOR, TER and GTM
- Human evaluation
- Remaining gaps

Automatic metrics of MT evaluation



Automatic metrics of MT evaluation

What can achieved with automatic evaluation
(as compared to manual evaluation)

- Automatic metrics notably accelerate the development cycle of MT systems:
  - Error analysis
  - System optimisation
  - System comparison

Besides, they are

- Costless (vs. costly)
- Objective (vs. subjective)
- Reusable (vs. non-reusable)

Automatic metrics of MT evaluation

Metrics based on lexical similarity
(most of the metrics!)

- Edit Distance: WER, PER, TER
- Precision: BLEU, NIST, WNM
- Recall: ROUGE, CDER
- Precision/Recall: GTM, METEOR, BLANC, SIA

Automatic metrics of MT evaluation

- Important Distinction:
  - Offline Evaluation of the MT system or Online Quality Assessment of MT performance?
  - **Main issue:** Do we have a pre-constructed sample set with target human reference translations to compare against?
- Reference-based Evaluation:
  - **Example:** Compare the performance of two MT systems on a sample set of client-specific documents
- "Reference-less" Confidence Scores:
  - **Example:** Filter out very poor MT translations so that they are not passed along to human post-editors

Automatic metrics of MT evaluation

Limits of lexical similarity

The reliability of lexical metrics depends very strongly on the heterogeneity/representativity of reference translations.

e: `This sentence is going to be difficult to evaluate.`

Ref1: `The evaluation of the translation is complicated.`
Ref2: `The sentence will be hard to qualify.`
Ref3: `The translation is going to be hard to evaluate.`
Ref4: `It will be difficult to punctuate the output.`

Lexical similarity is nor a sufficient neither a necessary condition so that two sentences convey the same meaning.

Automatic metrics of MT evaluation

Limits of lexical similarity

The reliability of lexical metrics depends very strongly on the heterogeneity/representativity of reference translations.

e: `This sentence is going to be difficult to evaluate.`

Ref1: `The evaluation of the translation is complicated.`
Ref2: `The sentence will be hard to qualify.`
Ref3: `The translation is going to be hard to evaluate.`
Ref4: `It will be difficult to punctuate the output.`

Lexical similarity is nor a sufficient neither a necessary condition so that two sentences convey the same meaning.

Automatic metrics of MT evaluation

- **Idea:** compare output of an MT system to a "reference" good (usually human) translation: how close is the MT output to the reference translation?
- **Advantages:**
  - Fast and cheap, minimal human labor, no need for bilingual speakers
  - Can be used on an on-going basis during system development to test changes
  - Minimum Error-rate Training (MERT) for search-based MT approaches!
- **Disadvantages:**
  - Current metrics are still relatively crude, do not distinguish well between subtle differences in systems
  - Individual sentence scores are often not very reliable, aggregate scores on a large test set are more stable
- Automated metrics for MT evaluation are still a very active area of current research

What do we want from MT metrics?

- High-levels of correlation with quantified human notions of translation quality
- Sensitive to small differences in MT quality between systems and versions of systems
- Consistent – same MT system on similar texts should produce similar scores
- Reliable – MT systems that score similarly will perform similarly
- General – applicable to a wide range of domains and scenarios
- Fast and lightweight – easy to run

Automated metrics for MT

- **Variety of Metric Uses and Applications:**
  - Compare (rank) performance of **different systems** on a common evaluation test set
  - Compare and analyze performance of different versions of **the same system**
    - Track system improvement over time
    - Which sentences got better or got worse?
  - Analyze the performance distribution of a **single system** across documents within a data set
  - Tune system parameters to optimize translation performance on a development set
- It would be nice if **one single metric** could do all of these well! But this is not an absolute necessity.
- A metric developed with one purpose in mind is likely to be used for other unintended purposes

History of automated metrics for MT

- 1990s: pre-SMT, limited use of metrics from speech – WER, PI-WER…
- 2002: IBM's BLEU Metric comes out
- 2002: NIST starts MT Eval series under DARPA TIDES program, using BLEU as the official metric
- 2003: Och and Ney propose MERT for MT based on BLEU
- 2004: METEOR first comes out
- 2006: TER is released, DARPA GALE program adopts HTER as its official metric
- 2006: NIST MT Eval starts reporting METEOR, TER and NIST scores in addition to BLEU, official metric is still BLEU
- 2007: Research on metrics takes off… several new metrics come out
- 2007: MT research papers increasingly report METEOR and TER scores in addition to BLEU
- 2008: NIST and WMT introduce first comparative evaluations of automatic MT evaluation metrics
- ….

Components of automated metrics for MT

- Example:
  - Reference: "the Iraqi weapons are to be handed over to the army within two weeks"
  - MT output: "in two weeks Iraq's weapons will give army"
- Possible metric components:
  - Precision: correct words / total words in MT output
  - Recall: correct words / total words in reference
  - Combination of P and R (i.e. F1= 2PR/(P+R))
  - Levenshtein edit distance: number of insertions, deletions, substitutions required to transform MT output to the reference
- Important Issues:
  - Features: matched words, ngrams, subsequences
  - Metric: a scoring framework that uses the features
  - Perfect word matches are weak features: synonyms, inflections: "Iraq's" vs. "Iraqi", "give" vs. "handed over"

BLEU

- Proposed by IBM [Papineni et al, 2002]
- Main ideas:
  - Exact matches of words
  - Match against a set of reference translations for greater variety of expressions
  - Account for Adequacy by looking at word precision
  - Account for Fluency by calculating n-gram precisions for n=1,2,3,4
  - No recall (because difficult with multiple refs)
  - To compensate for recall: introduce "Brevity Penalty"
  - Final score is weighted geometric average of the n-gram scores
  - Calculate aggregate score over a large test set
  - Not tunable to different target human measures or for different languages

Evaluation of MT quality

- Human and automatic MT evaluation: major issues
- **Automatic metrics**
  - **BLEU**
  - METEOR, TER and GTM
- Human evaluation
- Remaining gaps

BLEU

- Example:
  - Reference: "the Iraqi weapons are to be handed over to the army within two weeks"
  - MT output: "in two weeks Iraq's weapons will give army"
- BLUE metric:
  - 1-gram precision: 4/8
  - 2-gram precision: 1/7
  - 3-gram precision: 0/6
  - 4-gram precision: 0/5
  - BLEU score = 0 (weighted geometric average)

BLEU

- Clipping precision counts:
  - Reference1: "the Iraqi weapons are to be handed over to the army within two weeks"
  - Reference2: "the Iraqi weapons will be surrendered to the army in two weeks"
  - MT output: "the the the the"
  - Precision count for "the" should be "clipped" at two: max count of the word in any reference
  - Modified unigram score will be 2/4 (not 4/4)

BLEU

- Brevity Penalty:
  - Reference1: "the Iraqi weapons are to be handed over to the army within two weeks"
  - Reference2: "the Iraqi weapons will be surrendered to the army in two weeks"
  - MT output: "the Iraqi weapons will"
  - Precision score: 1-gram 4/4, 2-gram 3/3, 3-gram 2/2, 4-gram 1/1 → BLEU = 1.0
  - MT output is much too short, thus boosting precision, and BLEU doesn't have recall…
  - An exponential Brevity Penalty reduces score, calculated based on the aggregate length (not individual sentences)

BLEU

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

Then,

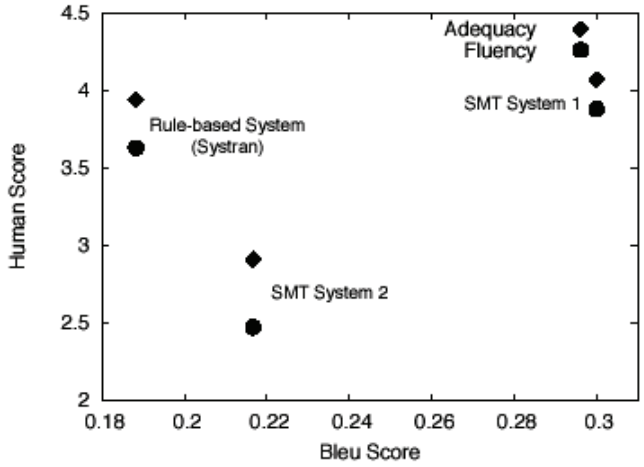
$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

$$\log \mathrm{BLEU} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n.$$

Weaknesses in BLEU

- BLEU matches word ngrams of MT-translation with multiple reference translations simultaneously → Precision-based metric
  - Is this better than matching with each reference translation separately and selecting the best match?
- BLEU Compensates for Recall by factoring in a "Brevity Penalty" (BP)
  - Is the BP adequate in compensating for lack of Recall?
- BLEU's ngram matching requires exact word matches
  - Can stemming and synonyms improve the similarity measure and improve correlation with human scores?
- All matched words weigh equally in BLEU
  - Can a scheme for weighing word contributions improve correlation with human scores?
- BLEU's higher order ngrams account for fluency and grammaticality, ngrams are geometrically averaged
  - Geometric ngram averaging is volatile to "zero" scores. Can we account for fluency/grammaticality via other means?

BLEU vs. Human evaluation



Evaluation of MT quality

- Human and automatic MT evaluation: major issues
- Automatic metrics
  - BLEU
  - **METEOR, TER and GTM**
- Human evaluation
- Remaining gaps

METEOR

- METEOR = Metric for Evaluation of Translation with Explicit Ordering [Lavie and Denkowski, 2009]
- Main ideas:
  - Combine Recall and Precision as weighted score components
  - Look only at unigram Precision and Recall
  - Align MT output with each reference individually and take score of best pairing
  - Matching takes into account translation variability via word inflection variations, synonymy and paraphrasing matches
  - Addresses fluency via a direct penalty for word order: how fragmented is the matching of the MT output with the reference?
  - Parameters of metric components **are tunable** to maximize the score correlations with human judgments
- METEOR has been shown to consistently outperform BLEU in correlation with human judgments

METEOR

- Unigram Precision: fraction of words in the MT that appear in the reference
- Unigram Recall: fraction of the words in the reference translation that appear in the MT
- F1= P*R/0.5*(P+R)
- Fmean = P*R/(α*P+(1-α)*R)
- Generalized Unigram matches:
  - Exact word matches, stems, synonyms
- Match with each reference separately and select the best match for each sentence

METEOR

- Example:
  - Reference: "the Iraqi weapons are to be handed over to the army within two weeks"
  - MT output: "in two weeks Iraq's weapons will give army"
- Matching: Ref: Iraqi weapons army two weeks
  MT: two weeks Iraq's weapons army
- P = 5/8 =0.625 R = 5/14 = 0.357
- Fmean = 10*P*R/(9P+R) = 0.3731
- Fragmentation: 3 frags of 5 words = (3-1)/(5-1) = 0.50
- Discounting factor: DF = 0.5 * (frag**3) = 0.0625
- Final score:

Fmean * (1- DF) = 0.3731 * 0.9375 = 0.3498

TER

- Translation Edit (Error) Rate, developed by Snover et. al. 2006
- Main Ideas:
  - Edit-based measure, similar in concept to Levenshtein distance: counts the number of word insertions, deletions and substitutions required to transform the MT output to the reference translation
  - Adds the notion of "block movements" as a single edit operation
  - Only exact word matches count, but latest version (TERp) incorporates synonymy and paraphrase matching and tunable parameters
  - Can be used as a rough post-editing measure
  - Serves as the basis for HTER – a partially automated measure that calculates TER between pre and post-edited MT output
  - Slow to run and often has a bias toward short MT translations

GTM

- General Text Matcher, developed by Turian et. al. 2003
- Main Ideas:
  - GTM measures similarity between the raw MT output (the "candidate" translation) and the reference sentence using measures of precision, recall and their composite F-measure (the harmonic mean).
  - Precision measures the number of words generated by the MT system that match with words in the reference sentence out of the total number of words generated by the MT system for that segment.
  - Recall measures the number of words generated by the MT system that match with words in the reference translation out of the total number of words in the reference translation.
  - The GTM metric also rewards matching adjacent words.
  - GTM shows better correlation with human judgments that other metrics (O'Brien, 2011).

MT confidence score

More information on Friday

Evaluation of MT quality

- Human and automatic MT evaluation: major issues
- Automatic metrics
  - BLEU
  - METEOR, TER and GTM
- **Human evaluation**
- Remaining gaps

Human evaluation

- Three main strategies:
  - Adequacy-fluency:
    - Fluency indicates how natural the hypothesis sounds to a native speaker of the target language.
    - Adequacy shows how much of the information from the original translation is expressed in the translation by selecting one of the proposed grades.
  - Ranking: annotators have to rank up to five sentences from best to worst relative to the other choices, with ties usually allowed.
  - Post-editing: annotators have to post-edit the references with information from the test hypothesis translations so that differences between a translation and reference account only for errors.
  - More informed: error classification

Human evaluation: ranking

Source: Estos tejidos están analizados, transformados y congelados antes de ser almacenados en Hema-Québec, que gestiona también el único banco público de sangre del cordón umbilical en Quebec.

Reference: These tissues are analyzed, processed and frozen before being stored at Héma-Québec, which manages also the only bank of placental blood in Quebec.

Translation	Rank
These weavings are analyzed, transformed and frozen before being stored in Hema-Quebec, that negotiates also the public only bank of blood of the umbilical cord in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
These tissues analysed, processed and before frozen of stored in Hema-Québec, which also operates the only public bank umbilical cord blood in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
These tissues are analyzed, processed and frozen before being stored in Hema-Québec, which also manages the only public bank umbilical cord blood in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
These tissues are analyzed, processed and frozen before being stored in Hema-Quebec, which also operates the only public bank of umbilical cord blood in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
These fabrics are analyzed, are transformed and are frozen before being stored in Hema-Québec, who manages also the only public bank of blood of the umbilical cord in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst

Human evaluation: ranking

EGMP-scale on sentence level (Roturier, 2009): a 4-level scale to measure output acceptability:

- Excellent (E): no post-editing required;
- Good (G): only minor post-editing is required;
- Medium (M): significant post-editing is required;
- Poor (P): it would be better to manually retranslate from scratch (post-editing is not worthwhile).

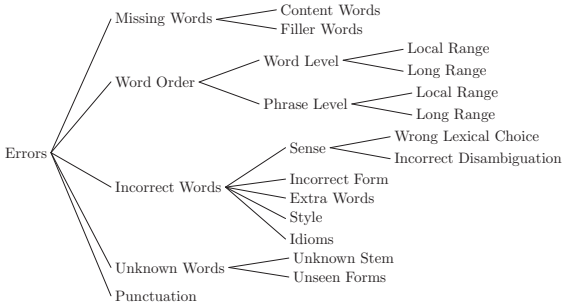
Human evaluation: H(metrics). Post-editing effort and adequacy/fluency.

- H(metrics): HTER, HBLEU, etc. manually construct reference translation for output, apply TER, BLEU or whatever have you (time consuming)
  - The National Institute of Standards for Technology (NIST) post-editing tool was
- Post-editing effort: ask the post-editors translate part of the sentences from scratch and post-edit the raw MT output for another part. Compare which activity is faster and for how much.

  But: time consuming, depend on skills of translator and post-editor.
- Adequacy/fluency evaluation:
  - Adequacy: Does the output convey the same meaning as the input sentence? Is part of the message lost, added, or distorted?
  - Fluency: Is the output good fluent target language (English)? This involves both grammatical correctness and idiomatic word choices.

Error classification

- Allows to perform error analysis of MT output (Vilar et al., 2006)



Evaluation of MT quality

- Human and automatic MT evaluation: major issues
- Automatic metrics
- BLEU
- METEOR, TER and GTM
- Human evaluation
- **Remaining gaps**

Remaining gaps

Recent efforts to go over lexical similarity

Extend the reference material:

- Using lexical variants such as morphological variations or synonymy lookup or using paraphrasing support.

Compare other linguistic features than words:

- Syntactic similarity: shallow parsing, full parsing (constituents /dependencies).
- Semantic similarity: named entities, semantic roles, discourse representations.

Combination of the existing metrics.

Remaining gaps

- Scores produced by most metrics are not intuitive or easy to interpret
- Scores produced at the individual segment-level are often not sufficiently reliable
- Need for greater focus on metrics with direct correlation with post-editing measures
- Need for more effective methods for mapping automatic scores to their corresponding levels of human measures (i.e. Adequacy)
- Need for more work on reference-less confidence scores for filtering poor MT (for post-editors and human translators)

Next session:

- Key problems that the SMT technology is facing

- Existing and foreseen solutions