Computational challenges for document translation

Sara Stymne
Dept. of Linguistics and Philology, UU

Joint work with Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre

2013-10-16

Research group

- Computational Linguistics group at Department of Linguistics and Philology, UU
- Main research areas
  - Statistical machine translation
  - Computational linguistics for the humanities
  - Syntactic parsing
- Common challenges
  - Learn about language from large text corpora

Statistical machine translation

- Data-driven
- Learn statistical models automatically from bilingual corpora
- Bilingual corpora: collections of texts translated by humans
- Use the models to translate unseen texts

Statistical machine translation – example

Amerikanska forskare säger sig ha funnit bevis för att det regnar diamanter på Saturnus och Jupiter. Metan faller från den övre delen av atmosfären, omvandlas till kolsot, och när regnet sedan faller mot marken blir det till grafit och till sist diamanter. Vad som händer sedan är lite svårare att säga, tycker forskarna. Men en möjlighet är att en slags ”sjö” av kol bildas.

(DN, Oct 15, 2013)

Statistical machine translation – example

Amerikanska forskare säger sig ha funnit bevis för att det regnar diamanter på Saturnus och Jupiter. Metan faller från den övre delen av atmosfären, omvandlas till kolsot, och när regnet sedan faller mot marken blir det till grafit och till sist diamanter. Vad som händer sedan är lite svårare att säga, tycker forskarna. Men en möjlighet är att en slags ”sjö” av kol bildas.

(DN, Oct 15, 2013)

U.S. scientists say they have found evidence that it is raining diamonds on Saturn and Jupiter. Methane falls from the upper part of the atmosphere, is converted to carbon soot, and when the rain fall to the ground then becomes to graphite and finally diamonds. What happens next is a bit harder to say, like scientists. But one possibility is that a kind of "lake" of the carbon formed.

(Google Translate, Oct 15, 2013)

Phrase-based SMT

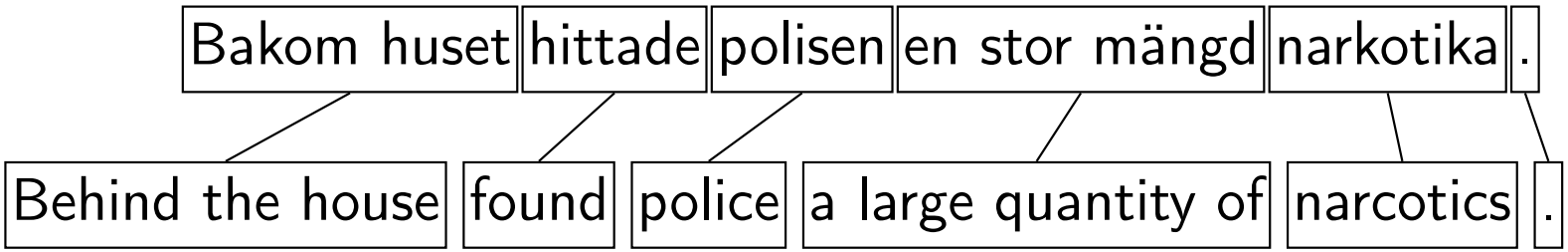
Bakom huset hittade polisen en stor mängd narkotika .

Phrase-based SMT

Bakom huset	hittade	polisen	en stor mängd	narkotika	.

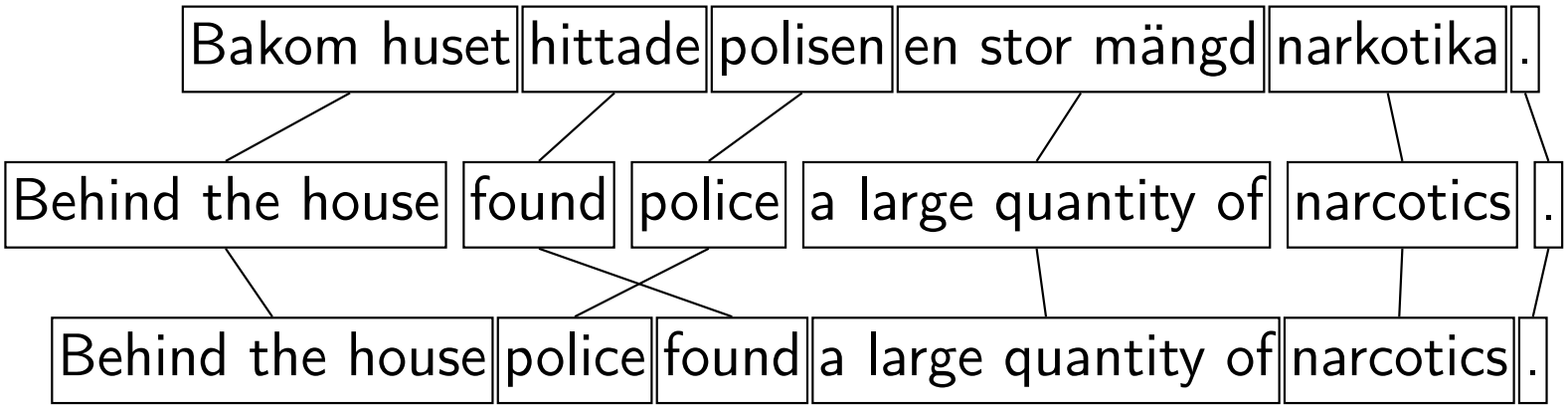
1. Phrase segmentation

Phrase-based SMT



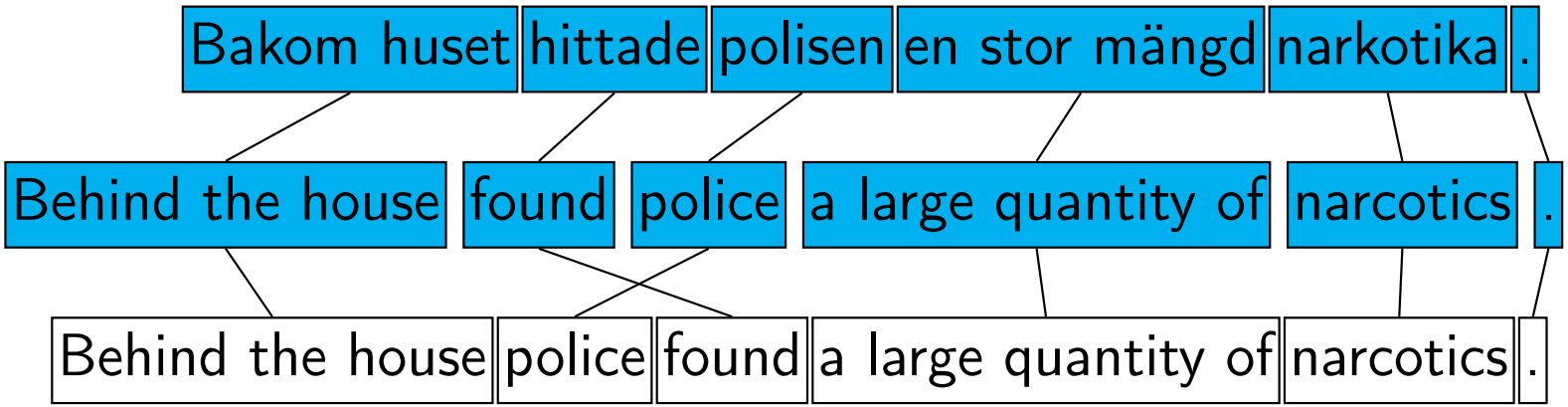
1. Phrase segmentation
2. Phrase translation

Phrase-based SMT



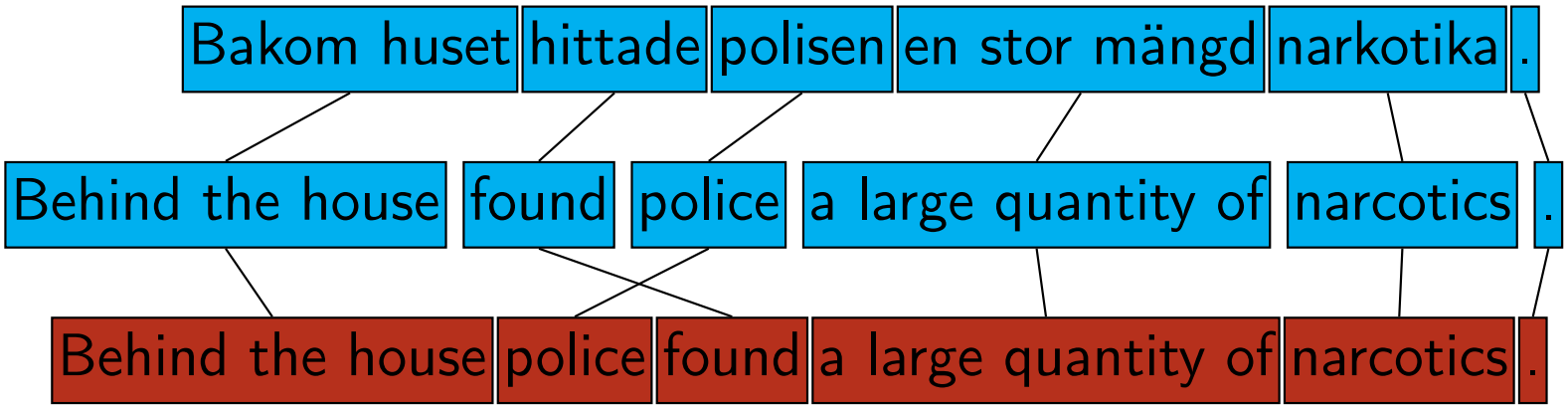
1. Phrase segmentation
2. Phrase translation
3. Output ordering

Phrase-based SMT



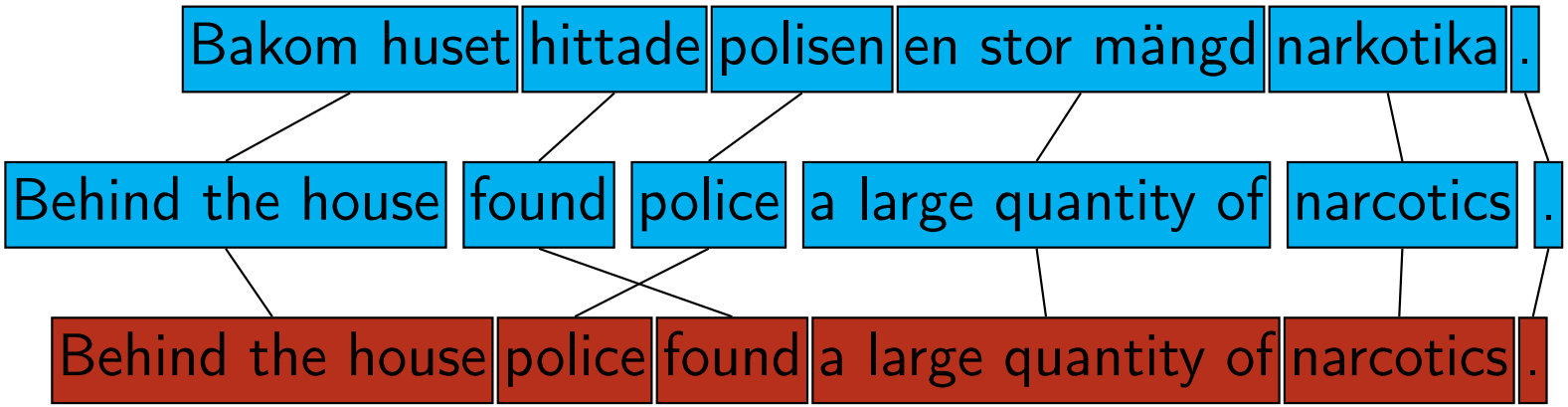
- Translation model

Phrase-based SMT



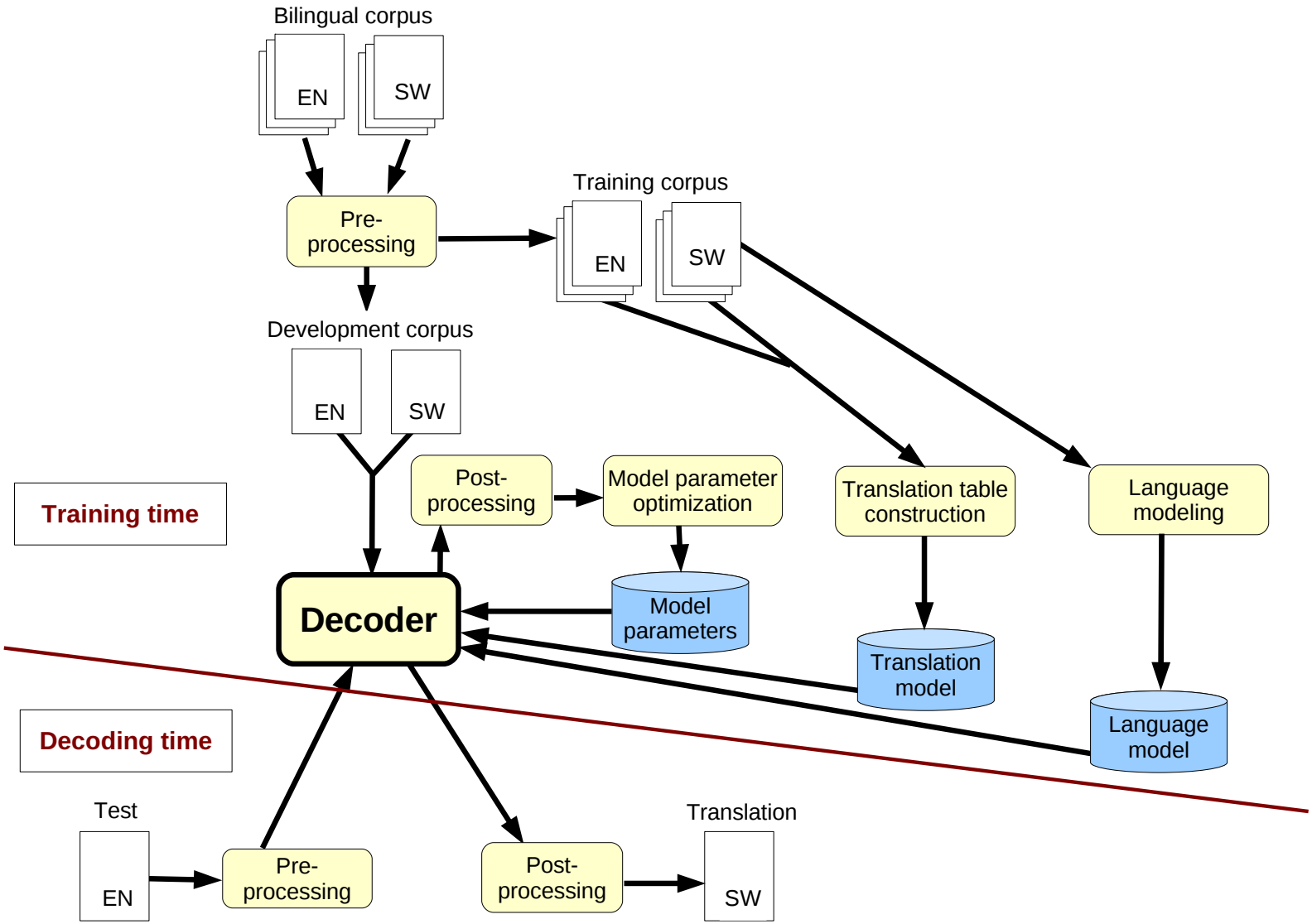
- Translation model
- Language model

Phrase-based SMT



- Translation model
- Language model
- Other models

Architecture



Language model

- How likely is a proposed target language sequence?
- Prefer grammatical/fluent strings
  - p(the house is small) $>$ p(small the is house)
  - p(he sleeps) $>$ p(he sleep)
  - p(small claim) $>$ p(little claim)
  - p(little girl) $>$ p(small girl)

Language model

- How likely is a proposed target language sequence?
- Prefer grammatical/fluent strings
  - p(the house is small) $>$ p(small the is house)
  - p(he sleeps) $>$ p(he sleep)
  - p(small claim) $>$ p(little claim)
  - p(little girl) $>$ p(small girl)
- N-gram models trained on large monolingual corpora

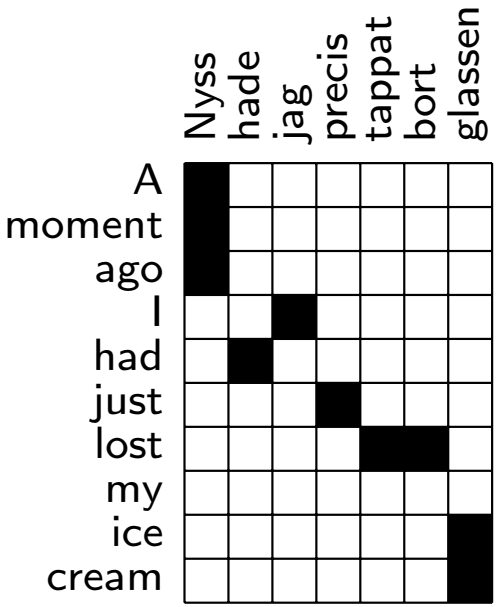
Translation model

- Phrase translations and their probabilities
- Example: phrase translations for **begreppet**

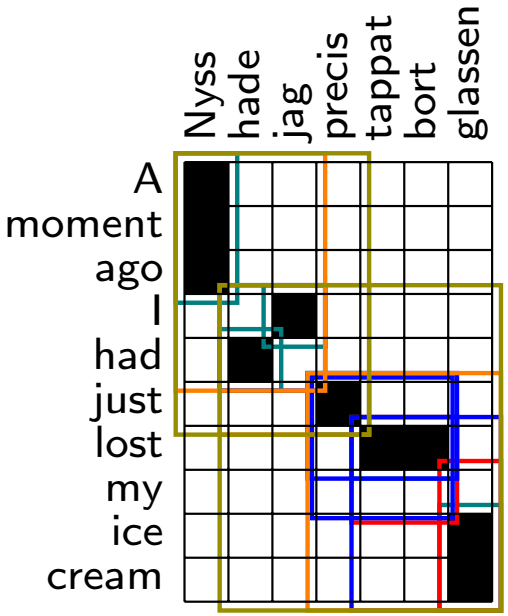
Target	**Probability** $\phi(\bar{t}\|\bar{s})$
announcement	0.075472
message	0.056604
information	0.037736
informed	0.037736
the information	0.008544
the information ,	0.005342
messages	0.001539
were told	0.000229
the back and	0.000003

Translation model learning

Align corresponding words



Translation model learning



Align corresponding words

Extract phrases:

a moment ago–nyss, I–jag, had–hade, just–precis
lost–tappat bort, ice cream–glassen

lost my–tappat bort, my ice cream–glassen

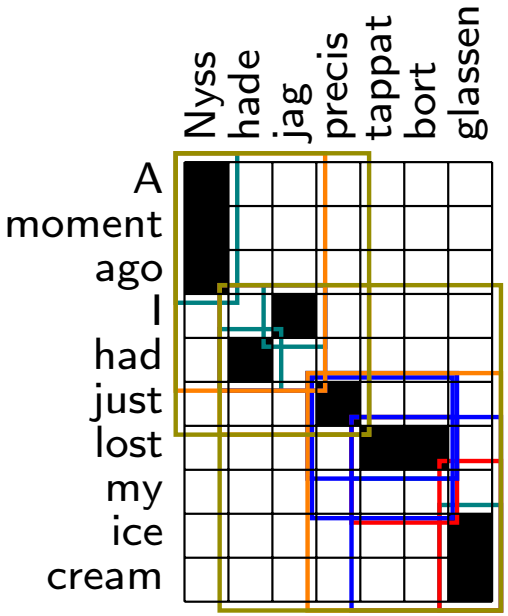
I had–hade jag, lost my ice cream–tappat bort glassen
just lost–precis tappat bort, just lost my–precis tappat bort

a moment ago I had–nyss hade jag, I had just–hade jag precis
just lost my ice cream–precis tappat bort glassen

a moment ago I had just–nyss hade jag precis
I had just lost my ice cream–hade jag precis tappat bort glassen

...

Translation model learning



Align corresponding words

Extract phrases:

a moment ago–nyss, I–jag, had–hade, just–precis
lost–tappat bort, ice cream–glassen

lost my–tappat bort, my ice cream–glassen

I had–hade jag, lost my ice cream–tappat bort glassen
just lost–precis tappat bort, just lost my–precis tappat bort

a moment ago I had–nyss hade jag, I had just–hade jag precis
just lost my ice cream–precis tappat bort glassen

a moment ago I had just–nyss hade jag precis
I had just lost my ice cream–hade jag precis tappat bort glassen

...

Score phrases based on relative frequency

Size of the phrase table

- Phrase translation table typically much larger than corpus
- Common to limit the length of phrase pairs (often to 7)
- Too big to store in memory?
  - Store on disk, read on demand
  - Use smart data structures, like suffix arrays
- Prune phrase table – i.e., remove non-useful phrase pairs
  - Limit translation options for each phrase (often to 20–30)
  - Prune table based on statistics, such as χ^2
- When new training data becomes available
  - Retrain the whole mode, or update incrementally?

Optimizing feature weights

- We use a log-linear model:

$$t^* = \arg\max_t \sum_i \lambda_i h_i(s,t)$$

  where the translation model, language model, and other models, h_i, are weighted
- How do we learn the best weights, λ_i?
- Optimize the weights on a small development corpus

Optimization

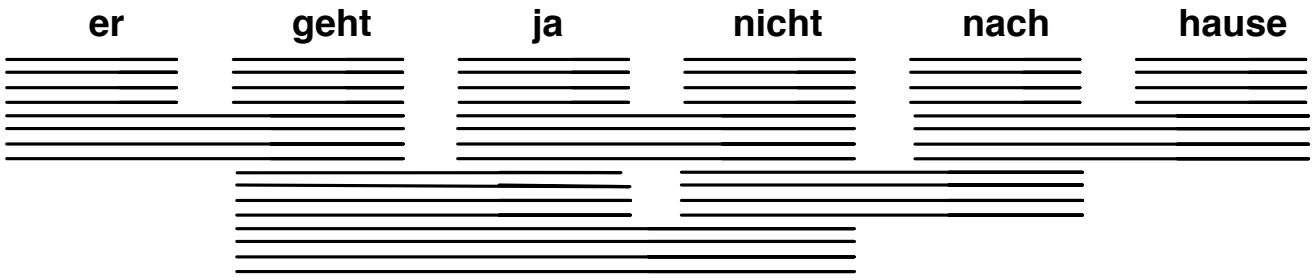
- Very hard optimization problem:
  - The correct output is unreachable
  - Latent variables in the translation process – non-convex loss functions
  - Sentence-level metrics are poor
- Many efficient standard optimization procedures cannot be used

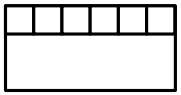
Decoding

- Decoding is the process of using all these models and weights to actually perform translation
- Find the best translation among all possible translations

$$t^* = \arg\max_t \sum_i \lambda_i h_i(s, t)$$

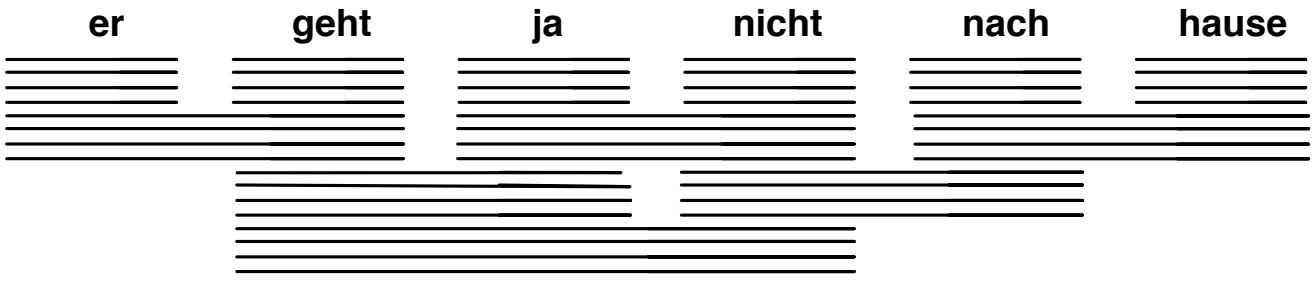
Decoding as search

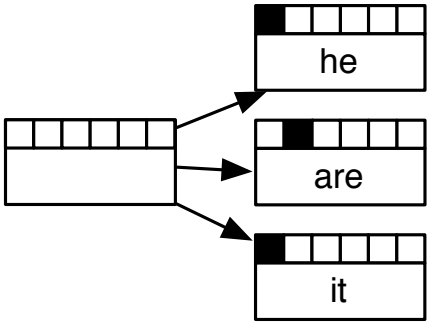




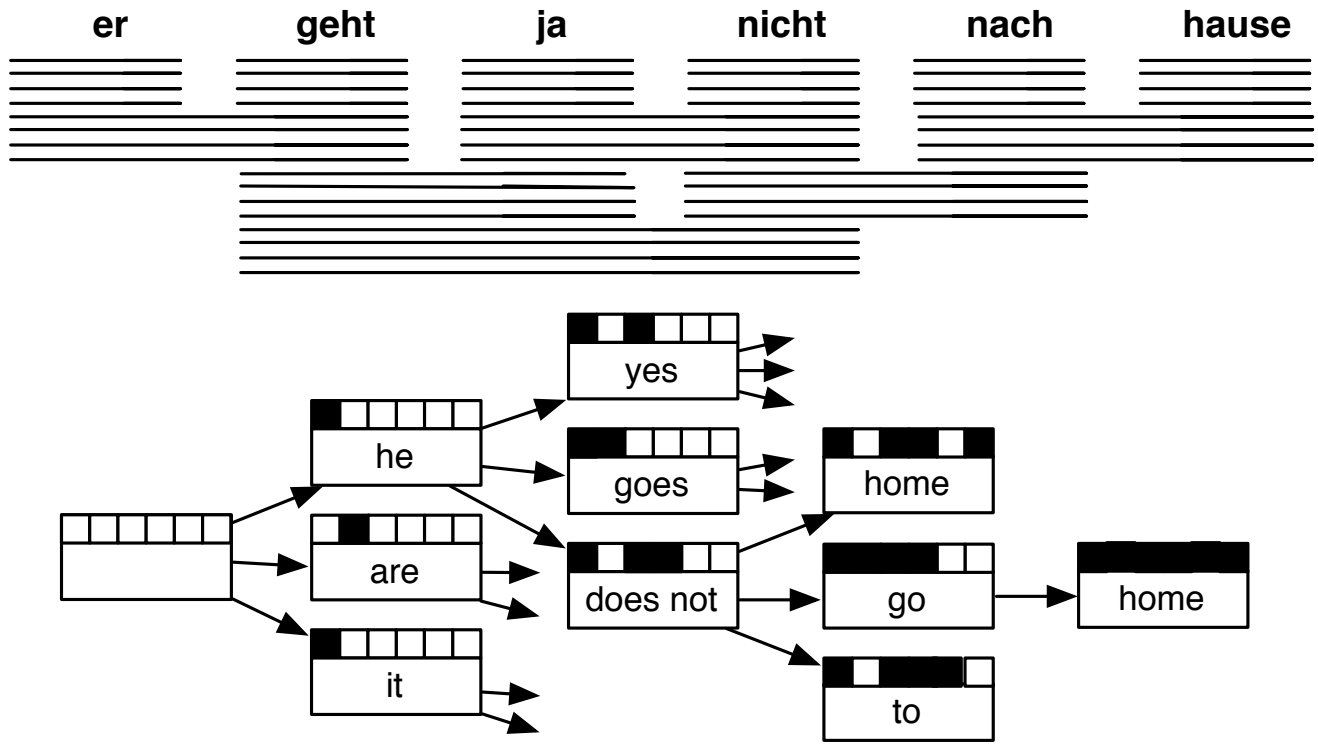
Illustrations by Philipp Koehn

Decoding as search





Decoding as search



Illustrations by Philipp Koehn

Decoding complexity

- Naively, in a sentence of N words with T translation options for each phrase, we can have
  - $O(2^N)$ phrase segmentations,
  - $O(T^N)$ sets of phrase translations, and
  - $O(N!)$ word reordering permutations.
- Unfeasible

Reduce decoding complexity

- Recombination – scoring is local, depends only on the last few words – recombine hypotheses that have an identical local history
- Prune – keep only S hypotheses for each word coverage
- Distortion limit – do not allow movement more than D words
- Reduces complexity to $O(N \cdot S)$ – but not optimal anymore

Document decoding

- The standard beam search decoding algorithm is fast and relatively accurate
  - Limited to very local models
- Our group is working on document-level decoding
  - We want to use non-local context
  - Model discourse phenomena
- The standard efficient models do not work anymore
- Scoring needs to access the full document

Docent – a document-wide decoder

- Hardmeier et al. (2013)
- A decoder based on local search
- Initializes a full translation for a document
- Uses hill-climbing to randomly pick an operation:
  - Change phrase translation
  - Resegment phrases
  - Move phrases
- Operations resulting in a better model score are kept

Issues with document decoding

- Stochastic search
  - No guarantees for finding the same translation
  - Empirically: surprisingly stable
- Scoring across documents is costly
  - Two step scoring, with only lower bounds for expensive models in the first step
- Optimization on document-level compared to sentence-level
  - Ongoing work
- Finding useful models
  - Hard to find models that improve state-of-the-art

Current document translation projects

- Pronoun anaphora
- Lexical consistency
- Translate into simplified Swedish
- Poetry translation

E-science related challenges for SMT

- Large amounts of data – that is also duplicated in models
- Hard search problems
- Hard optimization problems
- Parallel programming for speedups

Thank you

Tack

спасибо

Danke schön

Merci

Spurningar

Kesyon yo

Вопросы

Questions

Spørgsmål

Frågor

Domande

Vragen

Kysymyksiä

Otázky