UiO : University of Oslo

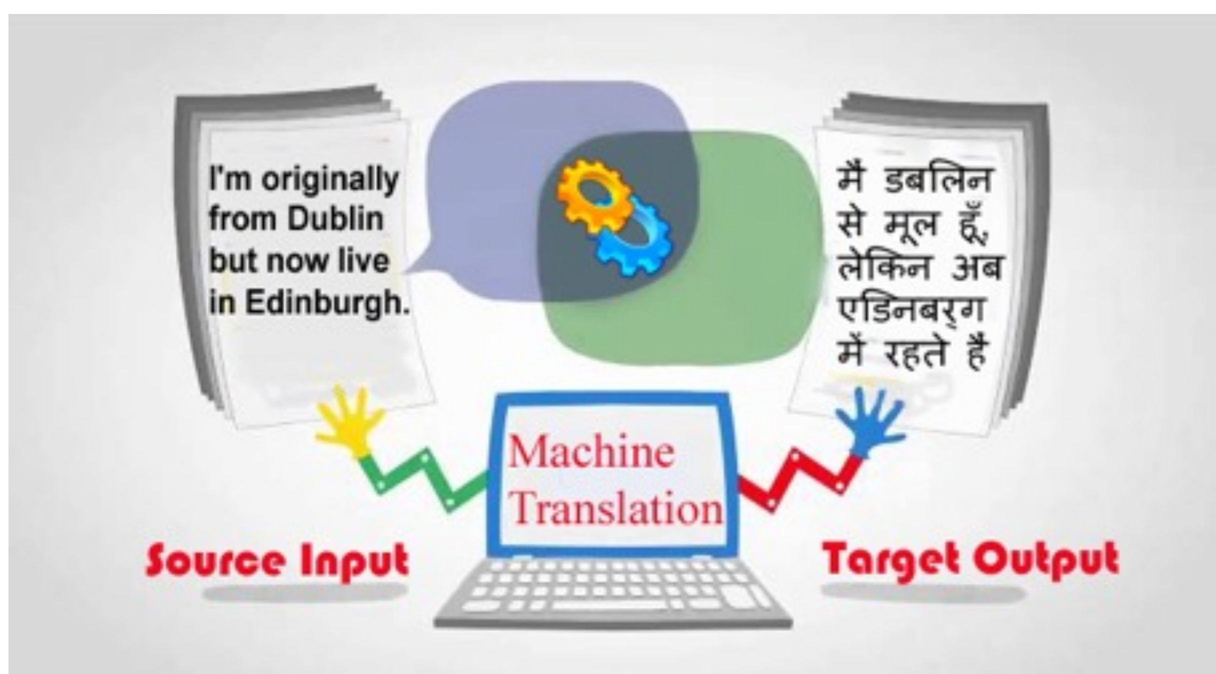
Introduction to Statistical Machine Translation

Pierre Lison
Language Technology Group
University of Oslo

Maskinlæring meetup, May 2016

Machine Translation?

- **Machine Translation** (MT) investigates how to automatically translate text or speech across (human) languages
  - Subfield of language technology / natural language processing
  - Long history in computer science, starting as early as 1949



Machine Translation?

- Hundreds of millions of users around the globe
    - Google Translate processes over **100 billion words a day**
- Most important use cases:
    - *Gisting*: Grasp the rough meaning of texts written in a foreign language
    - *Communicate* with others across language barriers
    - *Support* for human translation

Why should you care?

1. Complex, fascinating problem!
   - Infinite set of possible outputs
   - Sophisticated statistical models of linguistic structure
   - "AI-complete" problem!
2. MT can help *internationalisation* efforts
3. ... and make sense of *multilingual* data
4. *Useful insights* for other ML problems

Some challenges

- Ambiguities:

English: The *pen* was in the box **vs.** The box was in the *pen*

Norwegian: *Pennen* var i boksen — Boksen var i *bingen*

- Differences in word order:

German: Das rote Buch, das er auf den Tisch gelegt hat

French: Le livre rouge qu'il a mis sur la table

- Morphology (compounds, inflected forms, etc.):

Turkish: Avrupalılaştıramadıklarımızdanmışsınızcasına

English: As if you are reportedly of those of ours that we were unable to Europeanize

Outline

- Part 1: Key ideas of SMT systems

  I'll present here the key ideas behind (statistical) machine translation, such as translation models, language models and decoding.

- Part 2: Advanced topics

  I'll delve into more technical questions, such the extraction of word alignments from parallel data, the evaluation of machine translation systems, and some current "hot topics" in the field.

Machine translation approaches

	Rule-based MT	Statistical MT
	handcrafted rules to translate from source S to target T	*probabilistic models* $P(T\|S)$ estimated from parallel corpora
+	Fine-grained control over the translations	Robust, data-driven translation models
-	Expensive to build, limited coverage	Need large quantities of training data

Focus of this talk

Basic idea

Search for the most probable translation $\hat{T}$ for a given source sentence S:

$$\hat{T} = arg\,max_T \; P(T|S)$$
$$= arg\,max_T \; \frac{P(S|T)P(T)}{P(S)}$$
$$= arg\,max_T \; P(S|T)\,P(T)$$

(This is the "traditional" SMT model, we'll see later how to improve it)

Translation model
Encodes the faithfulness of T as a translation of S

Language model
Encodes the fluency of T in the target language

Translation model

- The first translation models relied on translation probabilities for individual words
- Did not account well for idiosyncratic expressions:

heavy → tung
smoker → røyker

but heavy smoker → ~~tung røyker~~

- **Better**: use translation tables for entire *phrases* instead

heavy	tung	0.95
heavy metal	heavy metal	0.61
heavy metal	tungmetal	0.34
smoker	røyker	0.99
heavy smoker	storrøyker	0.99
...	...	...

Note: a "phrase" can be here any sequence of words

How is this table derived? Wait until part 2 of this talk!

Translation model

S: À l'évidence, | son mari | était | un gros fumeur

T: Tydeligvis | var | mannen hennes | en storrøyker

We can then decompose the translation probability P(S|T) into I phrase pairs $\{(s_1,t_1),\ldots(s_I,t_I)\}$:

$$P(S|T) = \prod_{i=1}^{I} \phi(s_i|t_i)\, d(a_i - b_{i-1})$$

Phrase probability
(as given by the translation table)

Distortion probability
(relative distance between the phrase positions in the two languages)

Language model

- We also want the translated sentence T to be *fluent* in the target language
- A *statistical language model* is a probability distribution over sequences of words $w_1, w_2, \ldots w_n$
- Typically represented as N-grams:

$$P(w_1^n) = \prod_{k=1}^{n} P(w_k | w_1^{k-1})$$

Chain rule (the probability of each word depends on the words occurring before it)

$$\approx \prod_{k=1}^{n} P(w_k | w_{k-N+1}^{k-1})$$

Simplifying: we only consider the N previous words

word sequence $w_1, \ldots w_n$

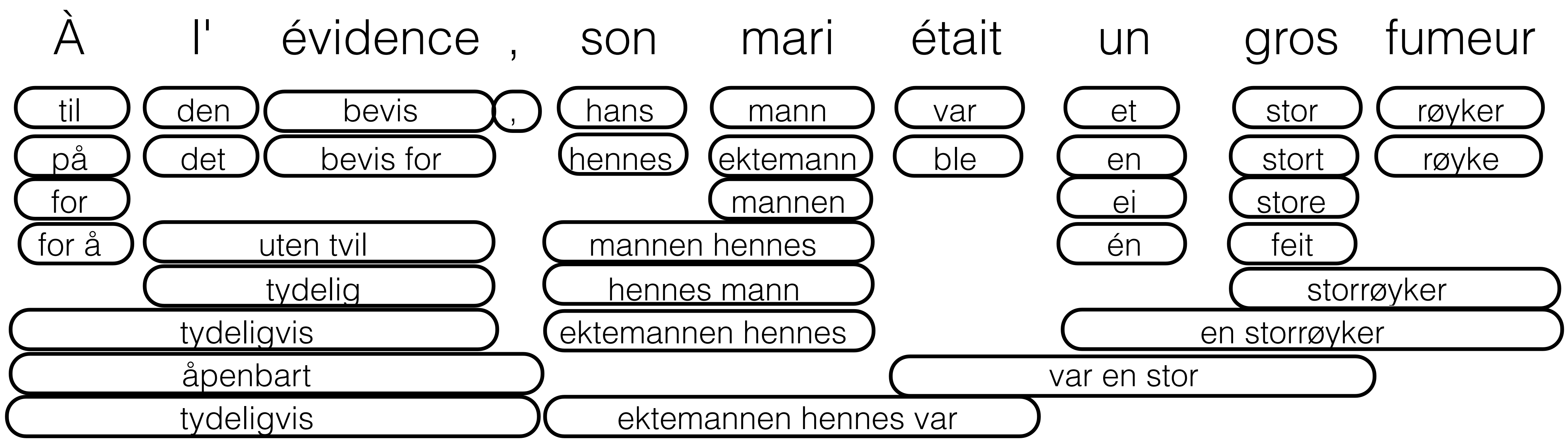
Language model

- The N-gram probabilities can be estimated from large amounts of monolingual data
  - Bigrams or trigrams are most popular
  - *Smoothing methods* to account for data sparsity
- Shortcoming: long-range dependencies

Das rote Buch, das <u>er</u> auf den Tisch gelegt <u>hat</u>

- New development: *neural language models* based on deep neural networks

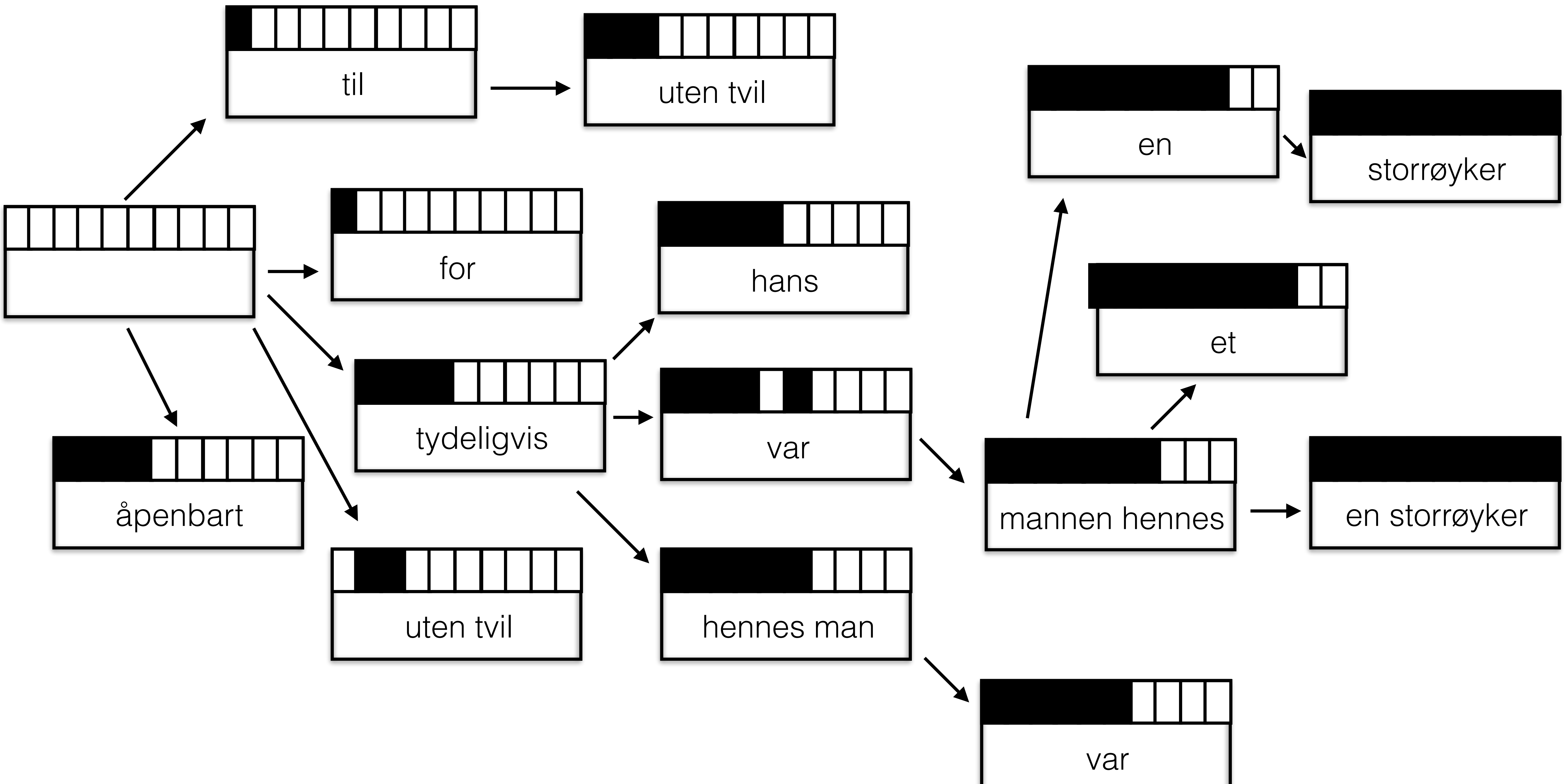
Decoding



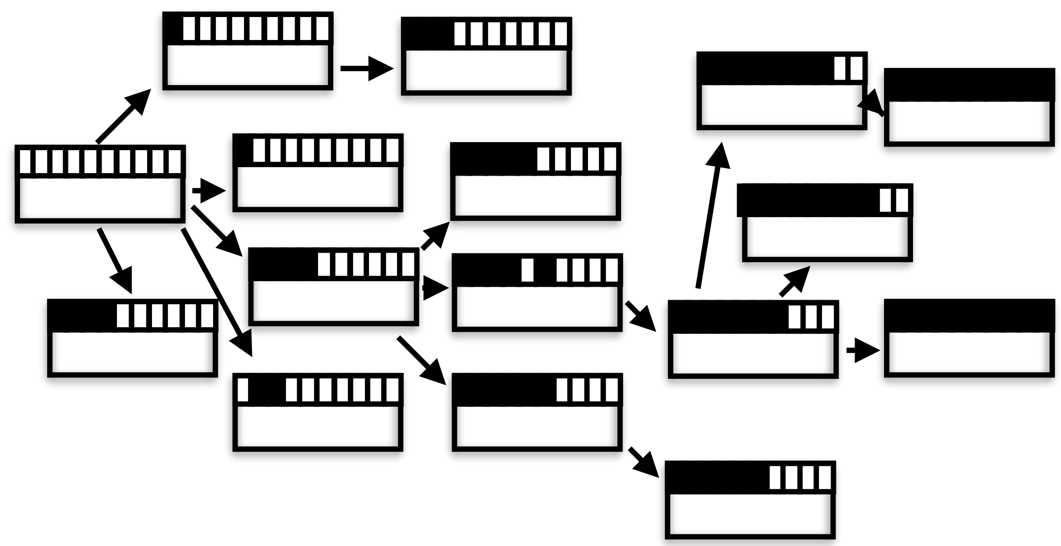
- How do we use the translation & language models to find the best translation *T* for a sentence *S*?
    - **Search** through the space of possible translations
- Incremental decoding process (**beam search**): gradual expansion of translation hypotheses

Decoding

À l' évidence , son mari était un gros fumeur



Decoding



- Every (partial) hypothesis is associated with a **cost**
  - Transition & language models + estimate of future costs
  - Beam search only keeps track of a limited number of good hypotheses (based on their cost), the rest is discarded
- The search space is further reduced through *hypothesis recombination*

What we've seen so far

- Translation as *probabilistic inference*: what is the most probable translation T for sentence S?
  - Based on a table with *possible phrase translations* ...
  - ... and a *language model* of the target language
  - *Decoding*: beam search for the best translation
- Still many open questions:
  - How do we *learn* this translation table from data?
  - How do we *evaluate* the quality of our translations?

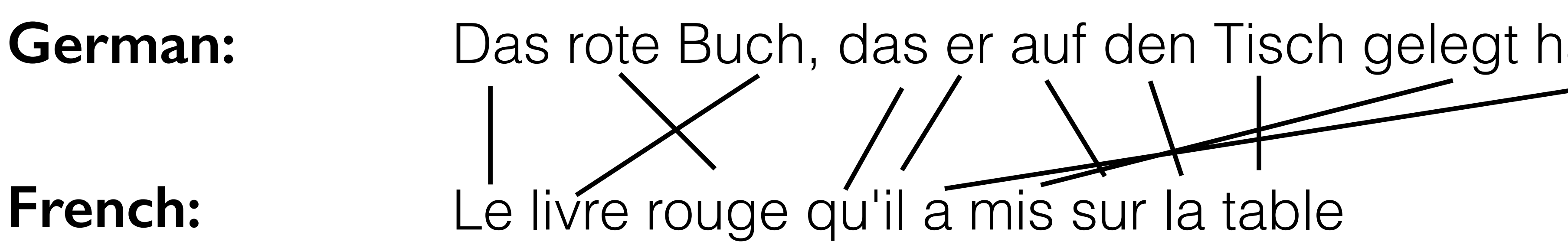
Parallel corpora

- Parallel corpora (or **bitexts**) are collections of texts available in (at least) two languages.
  - Alignment levels: documents, sentences, words
- Some examples:
  - Multilingual legal texts & parliament proceedings (EU, UN, etc.)
  - The Bible!
  - Translated sections of Wikipedia
  - Software localisation files
  - Movie subtitles

[Lison, P. & Tiedemann, J. (2016) OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. *LREC 2016]*

Alignment

- Parallel corpora typically aligned at sentence level
- ... But in order to extract pairs of phrase translations, we need *word alignments*

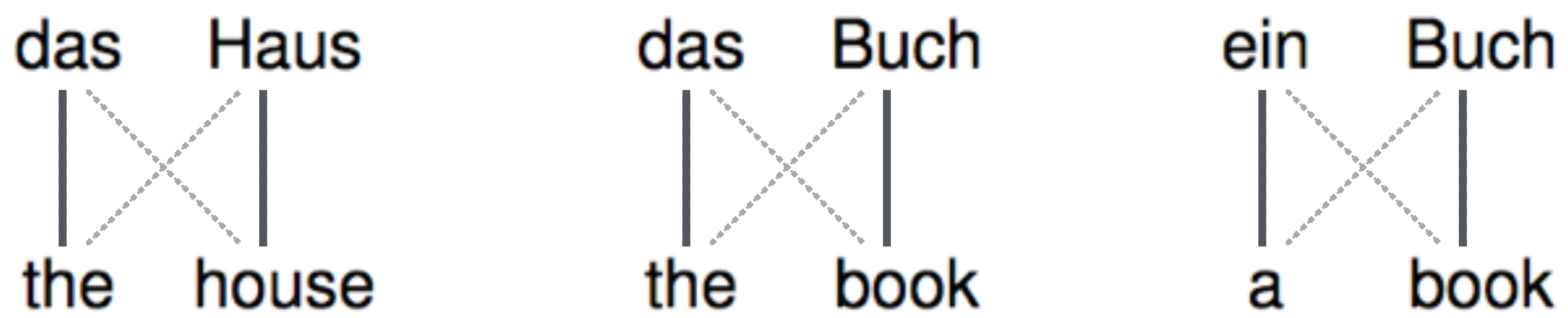


- Chicken-and-egg problem:
  - If we had a translation table, we could easily extract alignments
  - And if we had alignments, we could extract a translation table
  - But we have neither!

Alignment

- Solution: apply *Expectation-Maximisation* (EM)
  - The alignment is here the hidden variable
- Basic idea:
  - Start with uniform translation probabilities
  - Apply these probabilities to estimate possible alignments on the parallel sentences (**Expectation step**)
  - Revise the translation probabilities based on these alignments (**Maximisation step**)
  - Iterate until we have a stable solution

Example of alignment

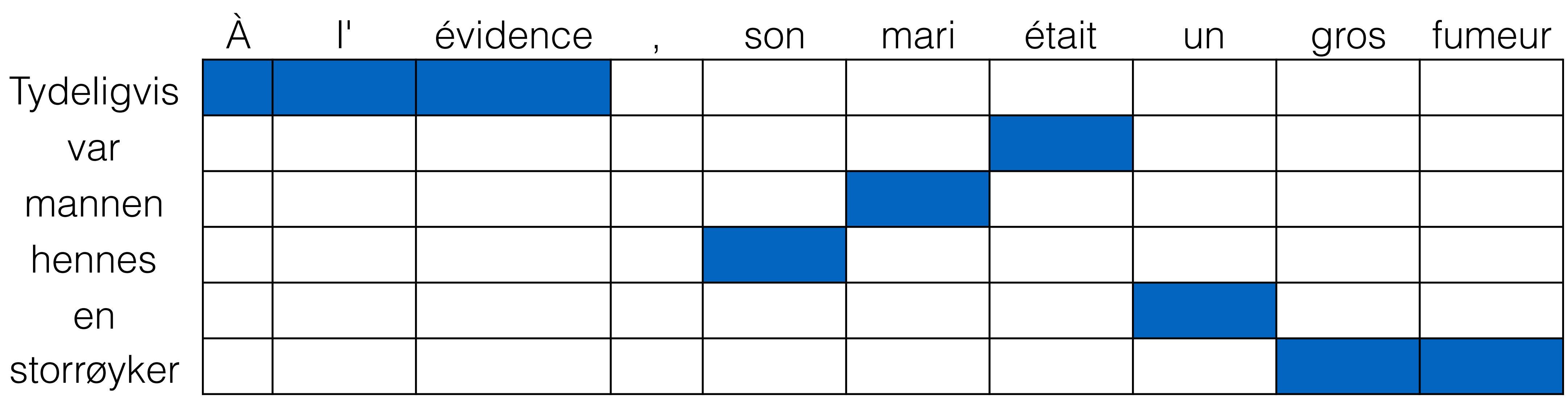


t	s	initial	1st it.	2nd it.	3rd it.	...	final
the	das	0.25	0.5	0.6364	0.7479	...	1
book	das	0.25	0.25	0.1818	0.1208	...	0
house	das	0.25	0.25	0.1818	0.1313	...	0
the	buch	0.25	0.25	0.1818	0.1208	...	0
book	buch	0.25	0.5	0.6364	0.7479	...	1
a	buch	0.25	0.25	0.1818	0.1313	...	0
book	ein	0.25	0.5	0.4286	0.3466	...	0
a	ein	0.25	0.5	0.5714	0.6534	...	1
the	haus	0.25	0.5	0.4286	0.3466	...	0
house	haus	0.25	0.5	0.5714	0.6534	...	1

(example borrowed from P. Koehn)

Phrase extraction

Finally, we can extract all phrase pairs that are consistent with the alignment:



tydeligvis	à l'évidence [,]
var	était
mannen	mari
hennes	[,] son
var mannen	mari était
mannen hennes	[,] son mari
var mannen hennes	[,] son mari était

en	un
storrøyker	gros fumeur
en storrøyker	un gros fumeur
var mannen hennes en storrøyker	[,] son mari était un gros fumeur
tydeligvis var mannen hennes en storrøyker	á l'évidence, son mari était un gros fumeur

Log-linear models

- Classical generative model seen so far:

$$\hat{T} = arg\,max_T \; P(S|T)\,P(T)$$

- Shortcomings:
  - The two models have fixed (equal) weights
  - Difficult to integrate other types of statistical models
- Modern MT systems adopt a *discriminative* approach where each model is seen as a feature function
  - Each model is also associated with a **weight**, which can be *tuned from data* to maximize *translation quality*

Log-linear models

$$\hat{T} = \arg\max_{T} P(T|S)$$

$$= \arg\max_{T} \exp\left[\sum_{m=1}^{M} \lambda_m h_m(T,S)\right]$$

(tunable) weight of model *m* | log-probability of (T,S) given *m*

In addition to the language model P(T) and translation model P(S|T), we can include other models such as:

- Reverse translation model P(T|S)
- Advanced reordering models
- Penalty scores to bias for shorter/longer translations

Evaluation

- Evaluation is of the most difficult problem in machine translation
  - What is a "good" translation, anyway?
  - Several alternative translations are often valid
- The ideal method is to rely on human raters to evaluate the translations
  - Key factors: **fluency** and **faithfulness**
  - But human evaluation is very expensive (and needs to be repeated each time the system is modified)

Evaluation

- Alternative: determine the translation quality based on its distance to some reference human translation(s)
- Most popular metric: **BLEU**
    - Based on *N-gram overlaps* with human translations

		Unigrams	Bigrams
Source:	À l'évidence, son mari était un gros fumeur		
Reference:	Tydeligvis var mannen hennes en storrøyker		
Output 1:	Mannen hennes var åpenbart en storrøyker	5	2
Output 2:	Hans ektemann var en tung røyker	2	0

Evaluation

- Main advantage of automatic metrics: can be done automatically!
  - Not perfect, but *correlated* with translation quality
- But they also have important shortcomings
  - Ignores the semantics of the sentence ("ikke" is just one word, but an important one!)
  - Ignores the global coherence/structure of the sentence
- Very active question in MT research

Some open questions

1. Dealing with complex morphology and syntax
   - How to integrate linguistic structure into the models?
   - Development of *factored* or *tree-based* approaches
2. Discourse aspects of translation
   - Current SMT systems operate *one sentence at a time*
   - Cross-sentential phenomena (e.g. coreference) are ignored
3. Deep neural networks?
   - Neural language models are now very successful
   - Now: End-to-end systems with *RNN encoder-decoders*

Want to build your own MT system?

- Check out **Moses:** http://statmt.org/moses/
  - Everything you need is the code and parallel data for your language pair(s)
  - Efficient beam search decoder
  - Various tools for preprocessing, training, tuning and testing
  - Actually documented!

MOSES
statistical
machine translation
system

Taking stock

- Statistical machine translation is now the dominant approach for MT today
  - Only need parallel data (and a good machine with lots of memory!) to translate between any language pair
- What if you have little or no data?
  - Handcrafted or hybrid MT systems may be more suited
- SMT methods can be used for other domains than classical translation!

Interested?

- MT is just one of many applications of language technology
- Other application domains:
  - Information extraction from text data
  - Speech recognition and synthesis
  - Conversational user interfaces

- If you want to know more (and maybe discuss future collaboration ideas?), feel free to contact me at **plison@ifi.uio.no**