COLIBRI: Constructions as Linguistic Bridges

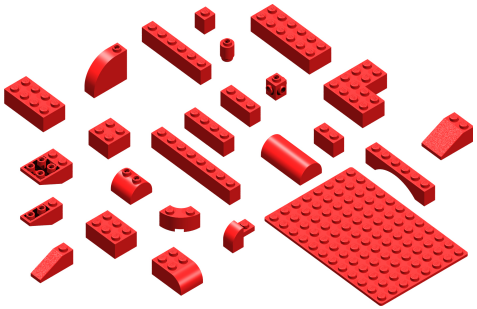
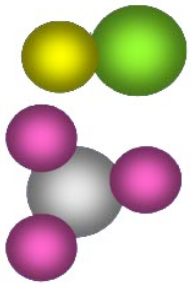
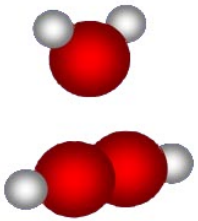
Maarten van Gompel, Radboud University Nijmegen

May 2012

What is a construction?

What is a construction?

- A *pattern* of words, *not necessarily consecutive*, which *in some way* forms an entity
- Constructions emerge from the data rather than linguistic theory: frequency threshold.
- Based on n-grams and *skipgrams*
- Intuitive "building blocks" for various NLP tasks
- A level above n-gram models and below syntactic level.

Skipgrams

- An n-gram with one or more gaps of specific length
- *"to be *2* to be"*
- *"ne *1* pas"*

Research focus

"Constructions" and their application in Machine Translation

Stages of research

1. Identification and extraction of constructions from monolingual corpus data
2. Alignment of constructions for a language pair (local translation step)
3. Machine learning to map constructions in context (local translation step)
4. Decoding (global translation step)

Pattern detection

Hypothesis

We can efficiently find constructions in large corpus data, limiting memory consumption

Counting n-grams

- Iterative counting, first unigrams, then bigrams,
- ... discarding pattern candidates if a sub-part is not found

Counting skip-grams

- Counting constructions with gaps: *"to be * to be"*
- ...by punching all combinations "holes" in found consecutive constructions
- ...discarding pattern candidates if consecutive sub-part is not found

Relations

Extracted patterns/constructions can be related in various ways and represented in a *graph model*

Motivations

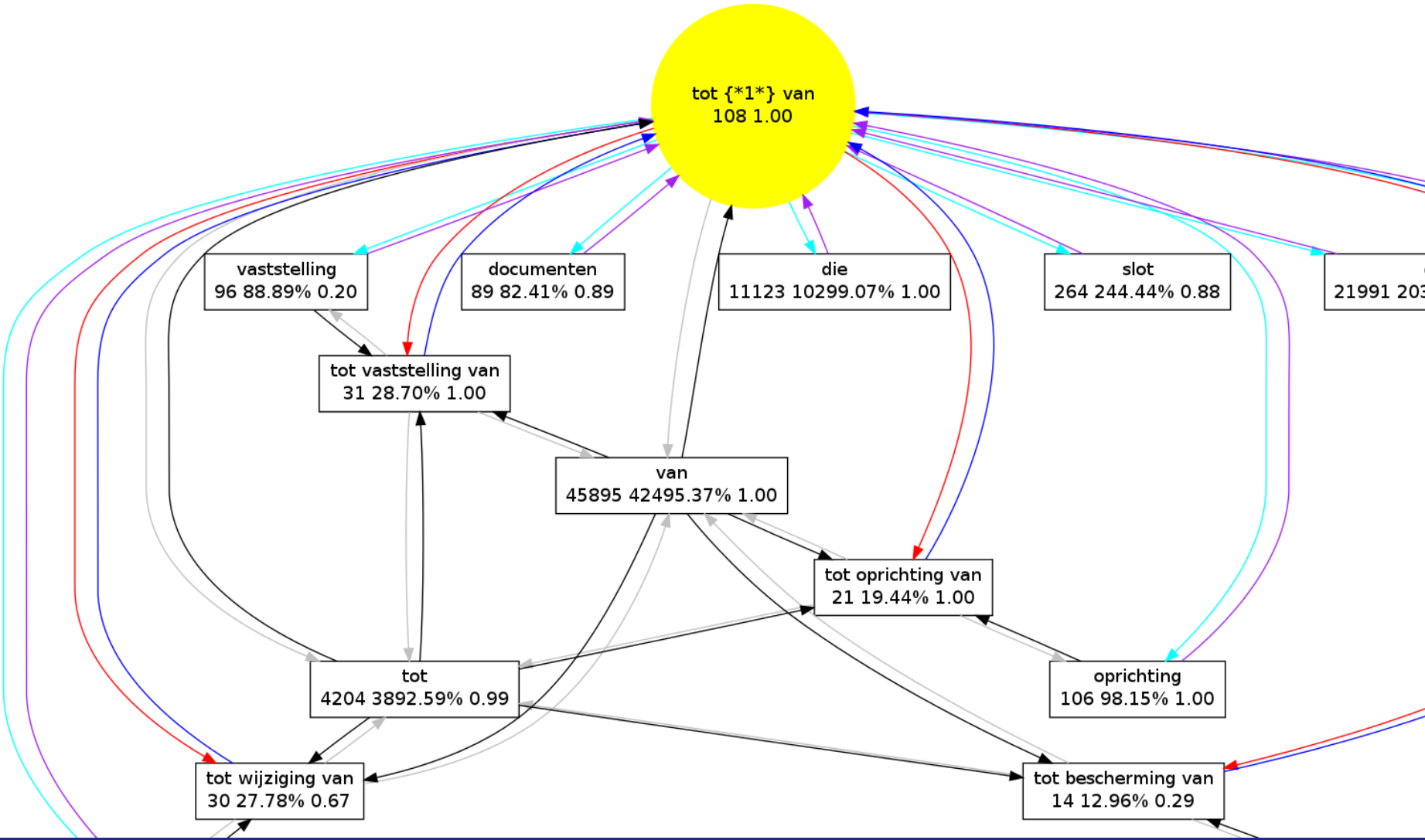
- To make explicit the information contained in the relations
- Extra information may help in constraining to "good" constructions and help obtain good alignments

Example

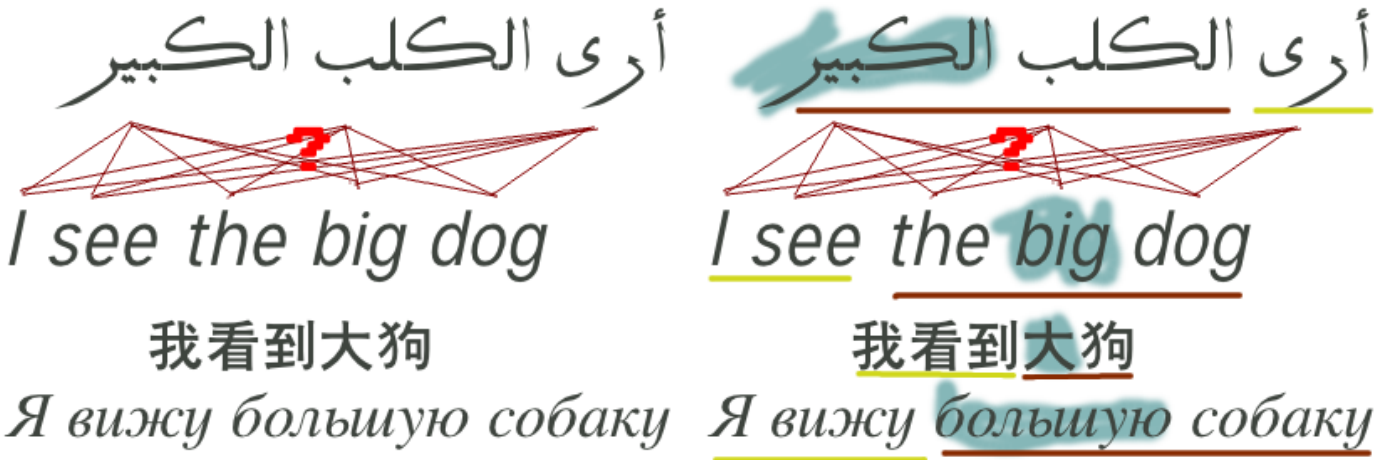
"I see the dog move"

- **Subsumption:** "I" is a sub-part of ''I see"
- **Succession:** "see" is a successor of "I"
- **Instantiation:** "I see the dog move" instance of "I see *2* move"

All relations: “tot ∗ van”



Alignment



Goal

Goal: phrase-translation table

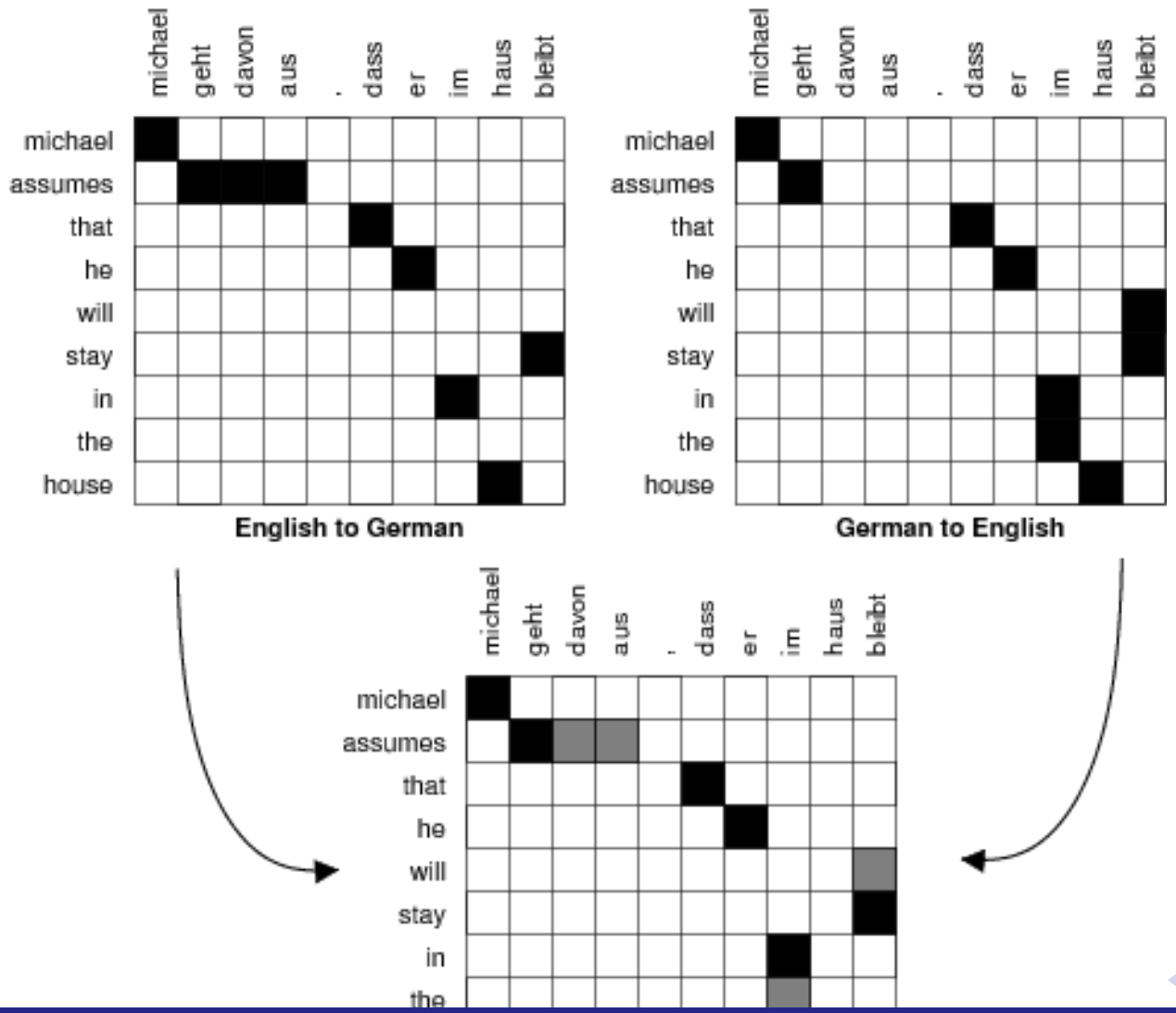
Question

Can we align extracted patterns directly?

Common method in Phrase-based Statistical MT

1. GIZA++ Word Alignment: source → target (IBM1, HMM, IBM4)
2. GIZA++ Word Alignment: target → source
3. Intersection of both
4. Heuristic methods adding certain alignment points from the union (grow-diag-final) (Och and Ney, 2003)
5. Extract all possible phrases

Alignment


michael geht davon aus . dass er im haus bleibt
michael
assumes
that
he
will
stay
in
the
house
English to German
German to English

Intuitions

- Word alignments are just an intermediate step towards phrasetables, we don't really need them.
- Word alignments introduce extra source for errors (poor quality alignments)
- Large memory availability might allow for more direct solutions

Our method

1. Extract constructions for source language
2. Extract constructions for target language
3. Attempt to directly align all constructions using EM (like IBM1):
   - source $\rightarrow$ target
   - target $\rightarrow$ source
   - intersection
4. Exploit graph-information (subsumption) to guide or adjust alignments

Graph information to aid alignment?

- ... by pre-pruning constructions to include only "exclusive" constructions
- ... by adjusting alignment weights (reward and punishment) based on subsumption relations

Difficulties

- Larger alignment matrix: larger memory consumption, scalability issues
- Competition of overlapping fragments? $\rightarrow$ skewed alignments

Preliminary results using Colibri alignment and Moses decoder (without skipgrams!)

- Evaluations scores (BLEU etc) a bit below the classic approach ($+-$ 0.01 BLEU-points)
- But: significantly smaller phrase-table $\rightarrow$ more generalisation
- Graph information not helpful in alignment yet

Future

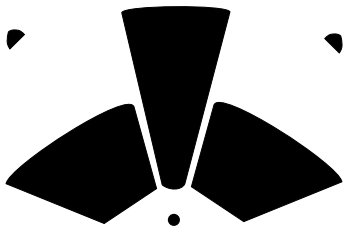
Future & Discussion

Alignment quality not sufficient yet..

- Reduce skewed alignments
- Scalability issues
- Graph information not helpful thus far, how to improve and integrate in EM?

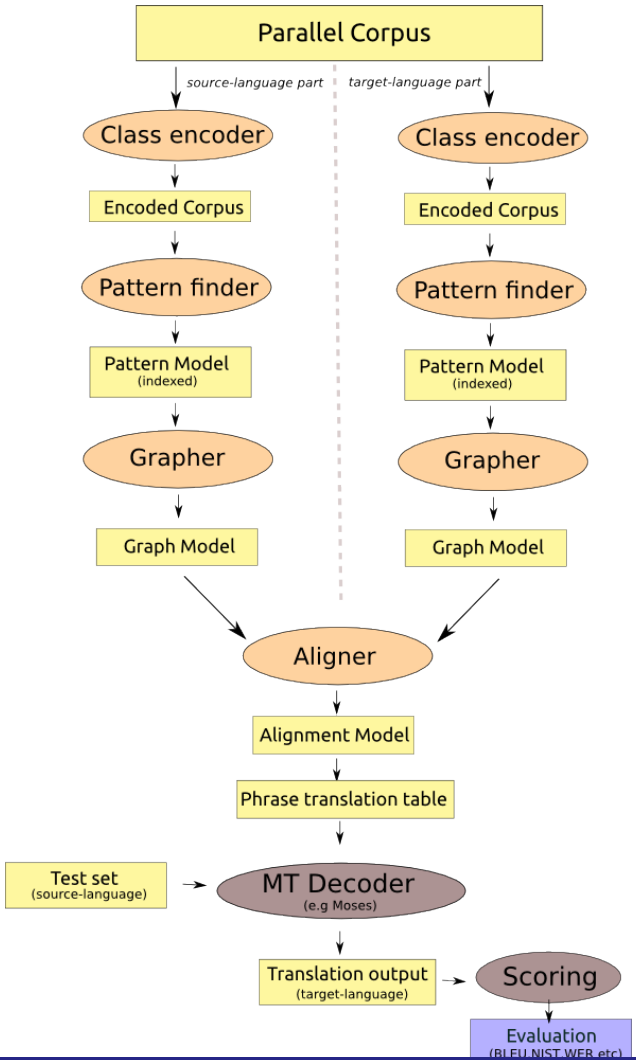
Shifting focus to skipgrams...

- Alignment of skipgrams
- Development of an MT decoder that supports skipgrams



Questions?

MT Pipeline


Parallel Corpus
source-language part
target-language part
Class encoder
Class encoder
Encoded Corpus
Encoded Corpus
Pattern finder
Pattern finder
Pattern Model
(indexed)
Pattern Model
(indexed)
Grapher
Grapher
Graph Model
Graph Model
Aligner
Alignment Model
Phrase translation table
Test set
(source-language)
MT Decoder
(e.g Moses)
Translation output
(target-language)
Scoring
Evaluation

IBM Model 1 EM

```
initialize t(t|s) uniformly
do until convergence
  set count(t|s) to 0 for all t,s
  set total(s) to 0 for all s
  for all sentence pairs (t_s,s_s)
     set total_s(t) = 0 for all t
     for all patterns t in t_s
        for all patterns s in s_s
          total_s(t) += t(t|s)
     for all patterns t in t_s
         for all patterns s in s_s
            count(t|s) += t(t|s) / total_s(t)
            total(s)   += t(t|s) / total_s(t)
  for all s
  for all t
       t(t|s) = count(t|s) / total(s)
```