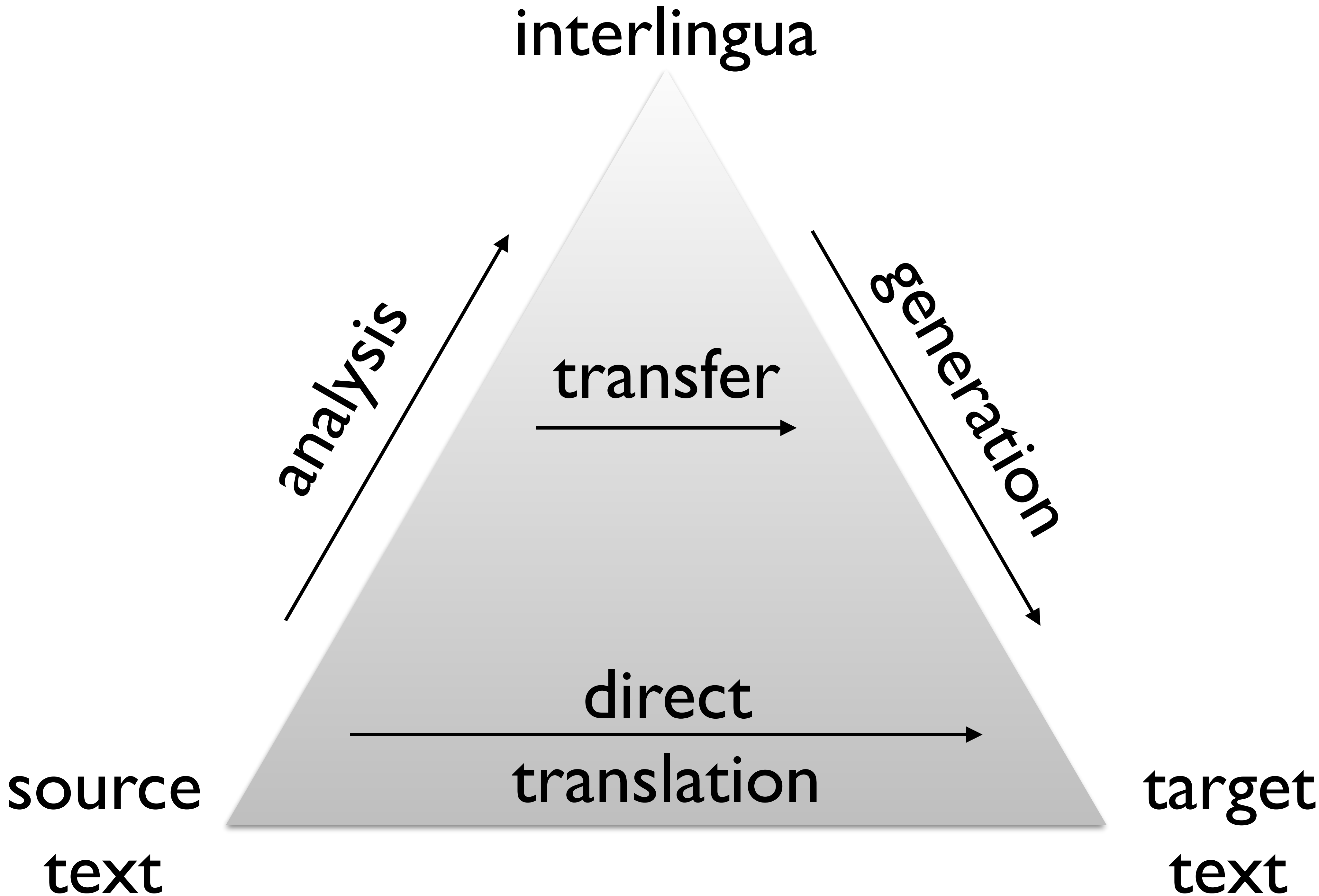
Machine Translation

Natural Language Processing
CS 4120/6120—Fall 2016
Northeastern University

David Smith
some slides from
Charles Schafer & Philip Koehn

Translation and NLP

- Translation is one of the oldest language tasks tried on a computer
  - Just look at that archaic name: “Machine Translation”!
- Translation involves many linguistic systems
- “Apollo program” dual-use argument:
  - Translation models of alignment and transfer are useful in question answering, paraphrase, information retrieval, etc.


interlingua
analysis
transfer
generation
direct
translation
source
text
target
text

Overview

- What problems does MT address? What does it (currently) not address?
- Models: What makes a good translation?
- Alignment: Learning dictionaries from parallel text
- Next: non-parallel text, translation decoding and training

The Translation Problem
and
Translation Data

The Translation Problem

মানব পরিবারের সকল সদস্যের সমান ও অবিচ্ছেদ্য অধিকারসমূহ এবং সহজাত মর্যাদার স্বীকৃতিই হচ্ছে বিশ্বে শান্তি, স্বাধীনতা এবং ন্যায়বিচারের ভিত্তি

The Translation Problem

মানব পরিবারের সকল সদস্যের সমান ও অবিচ্ছেদ্য অধিকারসমূহ এবং সহজাত মর্যাদার স্বীকৃতিই হচ্ছে বিশ্বে শান্তি, স্বাধীনতা এবং ন্যায়বিচারের ভিত্তি

The Translation Problem

মানব পরিবারের সকল সদস্যের সমান ও অবিচ্ছেদ্য অধিকারসমূহ এবং সহজাত মর্যাদার স্বীকৃতিই হচ্ছে বিশ্বে শান্তি, স্বাধীনতা এবং ন্যায়বিচারের ভিত্তি

Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world

Why Machine Translation?

* Cheap, universal access to world's online information regardless of original language. (That's the goal)

Why Statistical (or at least Empirical) Machine Translation?

* We want to translate real-world documents. Thus, we should model real-world documents.

* A nice property: design the system once, and extend to new languages automatically by training on existing data.

  F(training data, model) -> parameterized MT system

Ideas that cut across empirical language processing problems and methods

Real-world: don't be (too) prescriptive. Be able to process (translate/summarize/identify/paraphrase) relevant bits of human language as they are, not as they "should be". For instance, genre is important: translating French blogs into English is different from translating French novels into English.

Model: a fully described procedure, generally having variable parameters, that performs some interesting task (for example, translation).

Training data: a set of observed data instances which can be used to find good parameters for a model via a training procedure.

Training procedure: a method that takes observed data and refines the parameters of a model, such that the model is improved according to some objective function.

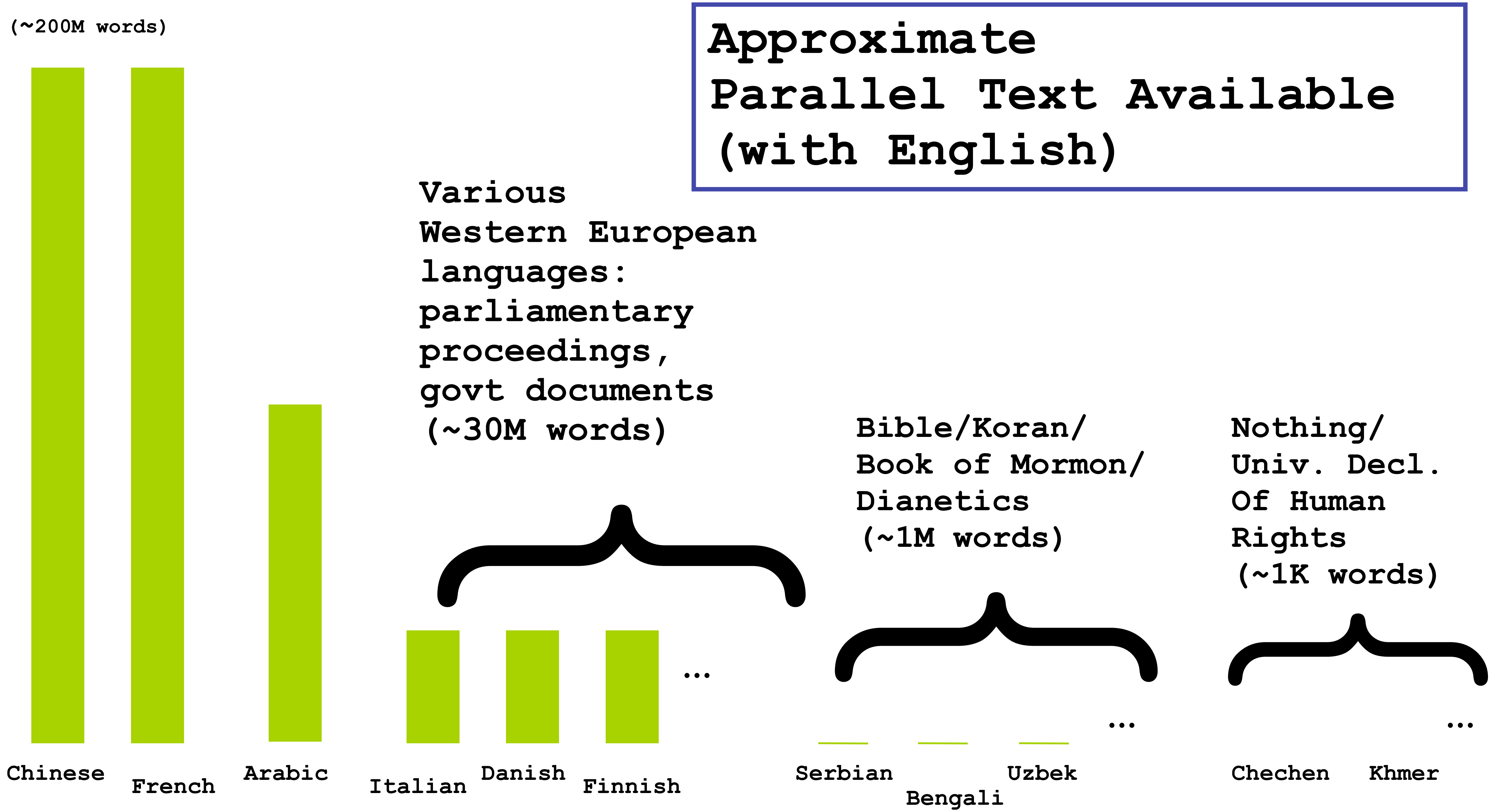
Resource Availability

Most of this lecture

Most statistical machine translation (SMT) research has focused on a few "high-resource" languages(European, Chinese, Japanese, Arabic).

Some other work: translation for the rest of the world's languages found on the web.

Most statistical machine translation research has focused on a few high-resource languages (European, Chinese, Japanese, Arabic).



Resource Availability

Most statistical machine translation (SMT) research has focused on a few "high-resource" languages(European, Chinese, Japanese, Arabic).

Some other work: translation for the rest of the world's languages found on the web.

Romanian Catalan Serbian Slovenian Macedonian Uzbek Turkmen Kyrgyz Uighur Pashto Tajikh Dari Kurdish Azeri Bengali Punjabi Gujarati Nepali Urdu Marathi Konkani Oriya Telugu Malayalam Kannada Cebuano

We'll discuss this briefly

<u>The Translation Problem</u>

<u>Document</u> translation? <u>Sentence</u> translation? <u>Word</u> translation?

What to translate? The most common use case is probably <u>document</u> translation.

Most MT work focuses on <u>sentence</u> translation.

What does sentence translation ignore?

- **Discourse properties/structure.**
- **Inter-sentence coreference.**

Sentence Translation

- SMT has generally ignored extra-sentence structure (good future work direction for the community).

- Instead, we've concentrated on translating individual sentences as well as possible. This is a very hard problem in itself.

- Word translation (knowing the possible English translations of a French word) is not, by itself, sufficient for building readable/useful automatic document translations – though it is an important component in end-to-end SMT systems.

Sentence translation using only a word translation dictionary is called "glossing" or "gisting".

Word Translation (learning from minimal resources)

We'll come back to this later…

and address learning the word translation component (dictionary) of MT systems without using parallel text.

(For languages having little parallel text, this is the best we can do right now)

Sentence Translation

- Training resource: parallel text (bitext).

- Parallel text (with English) on the order of 20M-200M words (roughly, 1M-10M sentences) is available for a number of languages.

- Parallel text is expensive to generate: human translators are expensive ($0.05-$0.25 per word). Millions of words training data needed for high quality SMT results. So we take what is available. This is often of less than optimal genre (laws, parliamentary proceedings, religious texts).

Sentence Translation: examples of more and less literal translations in bitext

French, English from Bitext	Closely Literal English Translation
Le débat est clos . **The debate is closed .**	*The debate is closed.*
Accepteriez - vous ce principe ? **Would you accept that principle ?**	*Accept-you that principle?*
Merci , chère collègue . **Thank you , Mrs Marinucci .**	*Thank you, dear colleague.*
Avez - vous donc une autre proposition ? **Can you explain ?**	*Have you therefore another proposal?*

(from French-English European Parliament proceedings)

Sentence Translation: examples of more and less literal translations in bitext

Word alignments illustrated.
Well-defined for more literal translations.

Le débat est clos .

The debate is closed .

Accepteriez - vous ce principe ?

Would you accept that principle ?

Merci , chère collègue .

Thank you , Mrs Marinucci .

Avez - vous donc une autre proposition ?

Can you explain ?

Translation and Alignment

- As mentioned, translations are expensive to commission and generally SMT research relies on already existing translations

- These typically come in the form of aligned documents.

- A sentence alignment, using pre-existing document boundaries, is performed automatically. Low-scoring or non-one-to-one sentence alignments are discarded. The resulting aligned sentences constitute the training bitext.

- For many modern SMT systems, induction of word alignments between aligned sentences, using algorithms based on the IBM word-based translation models, is one of the first stages of processing. Such induced word alignments are generally treated as part of the observed data and are used to extract aligned phrases or subtrees.

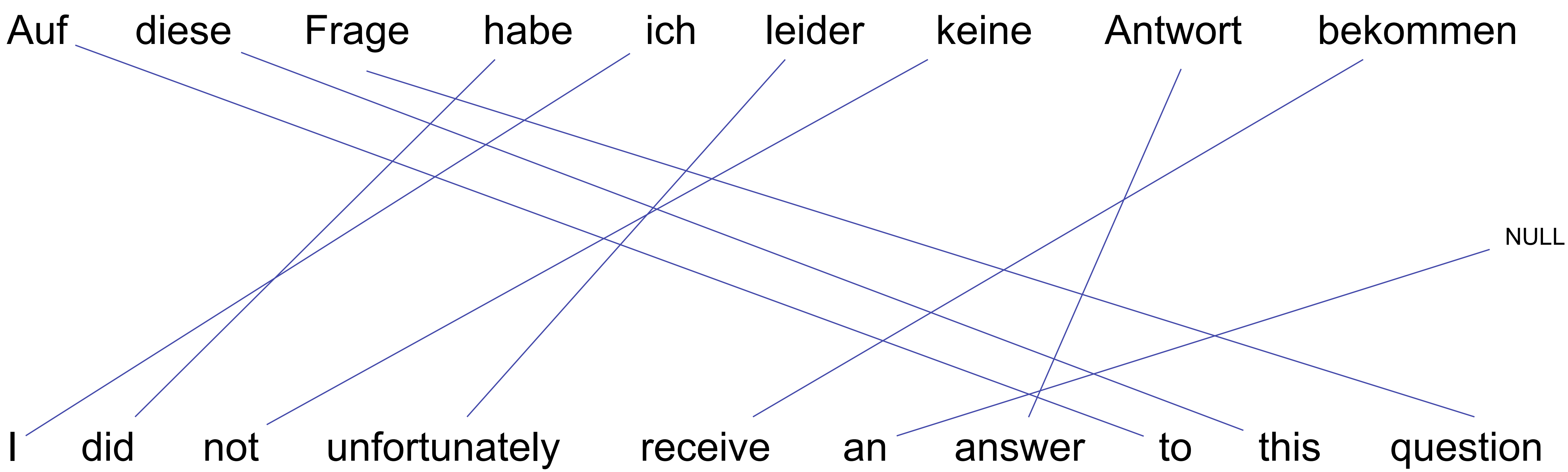
Modeling

What Makes a Good Translation?

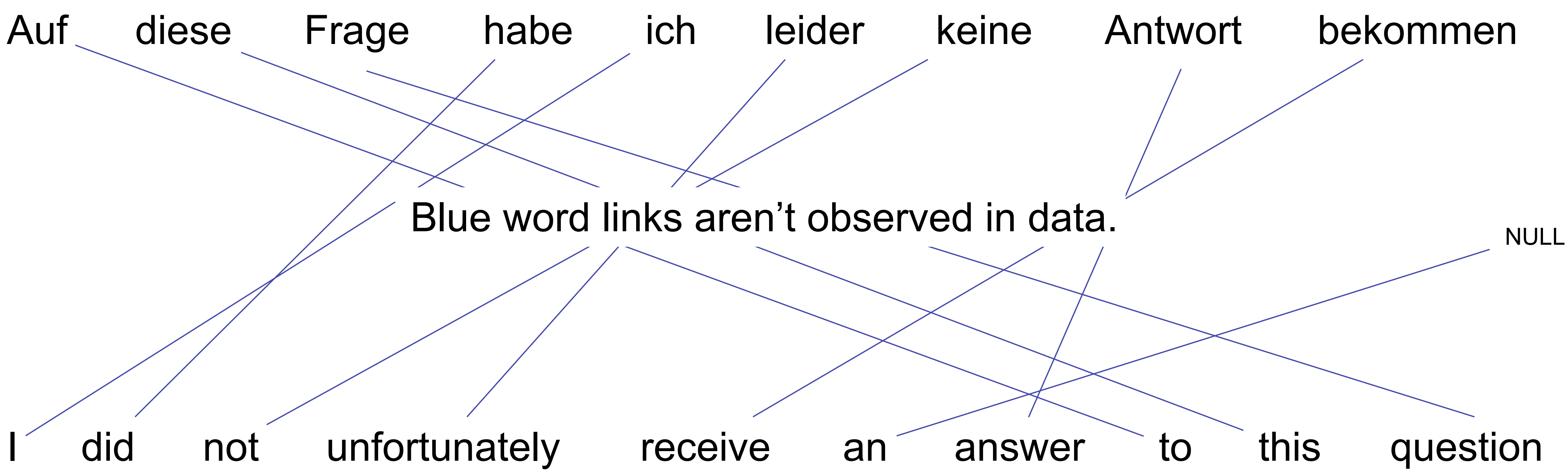
Modeling

- Translation models
  - “Adequacy”
  - Assign better scores to accurate (and complete) translations
- Language models
  - “Fluency”
  - Assign better scores to natural target language text

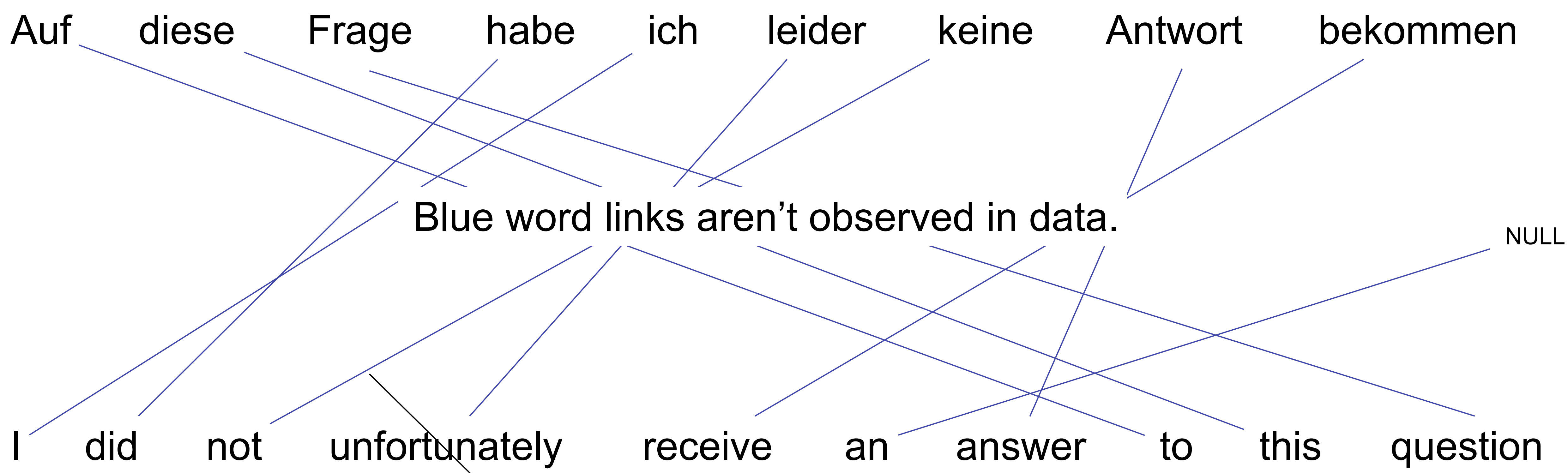
Word Translation Models



Word Translation Models



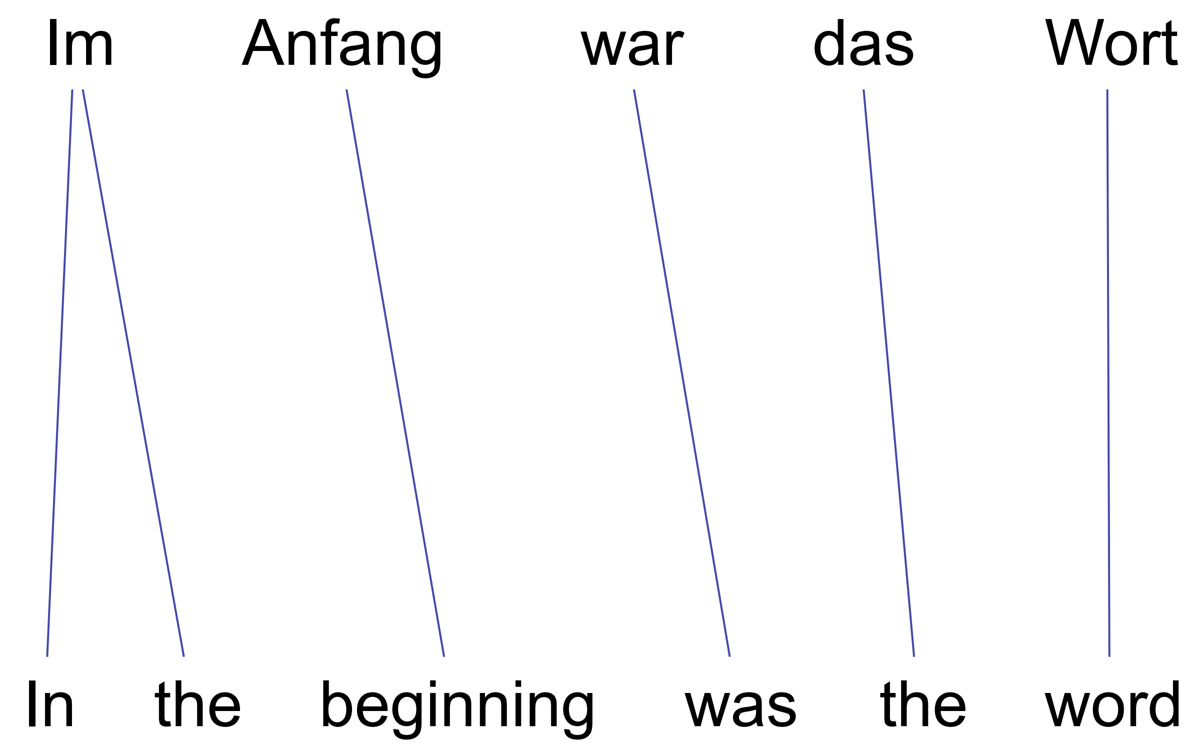
Word Translation Models



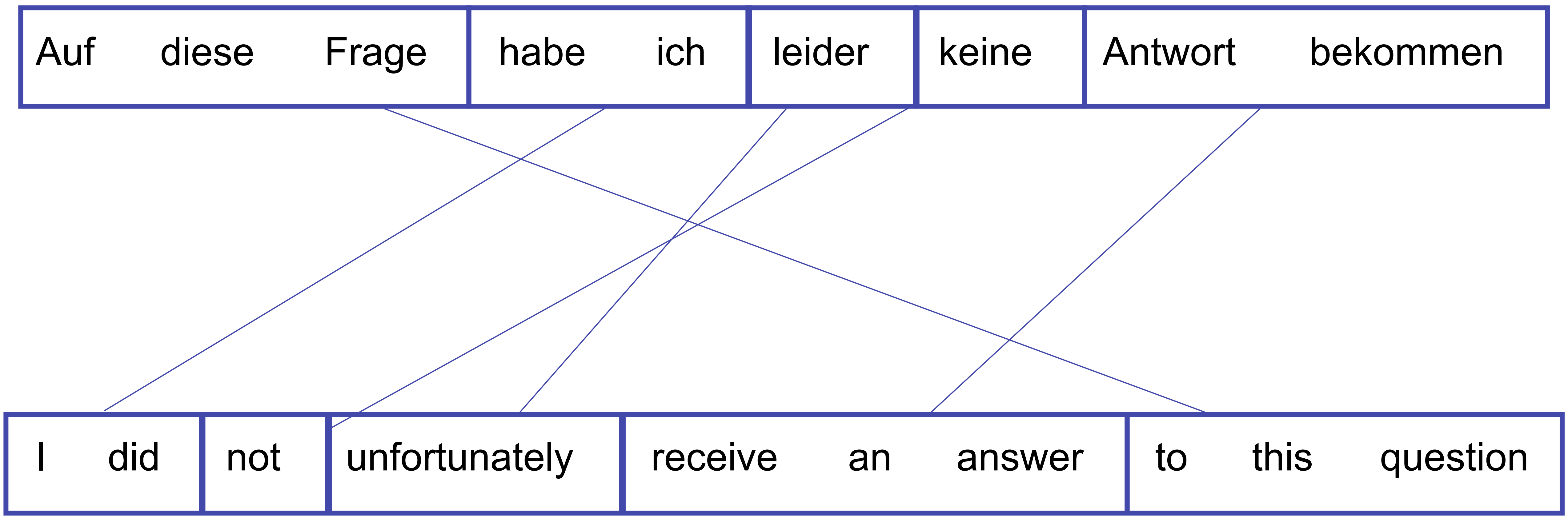
Features for word-word links: lexica, part-of-speech, orthography, etc.

Word Translation Models

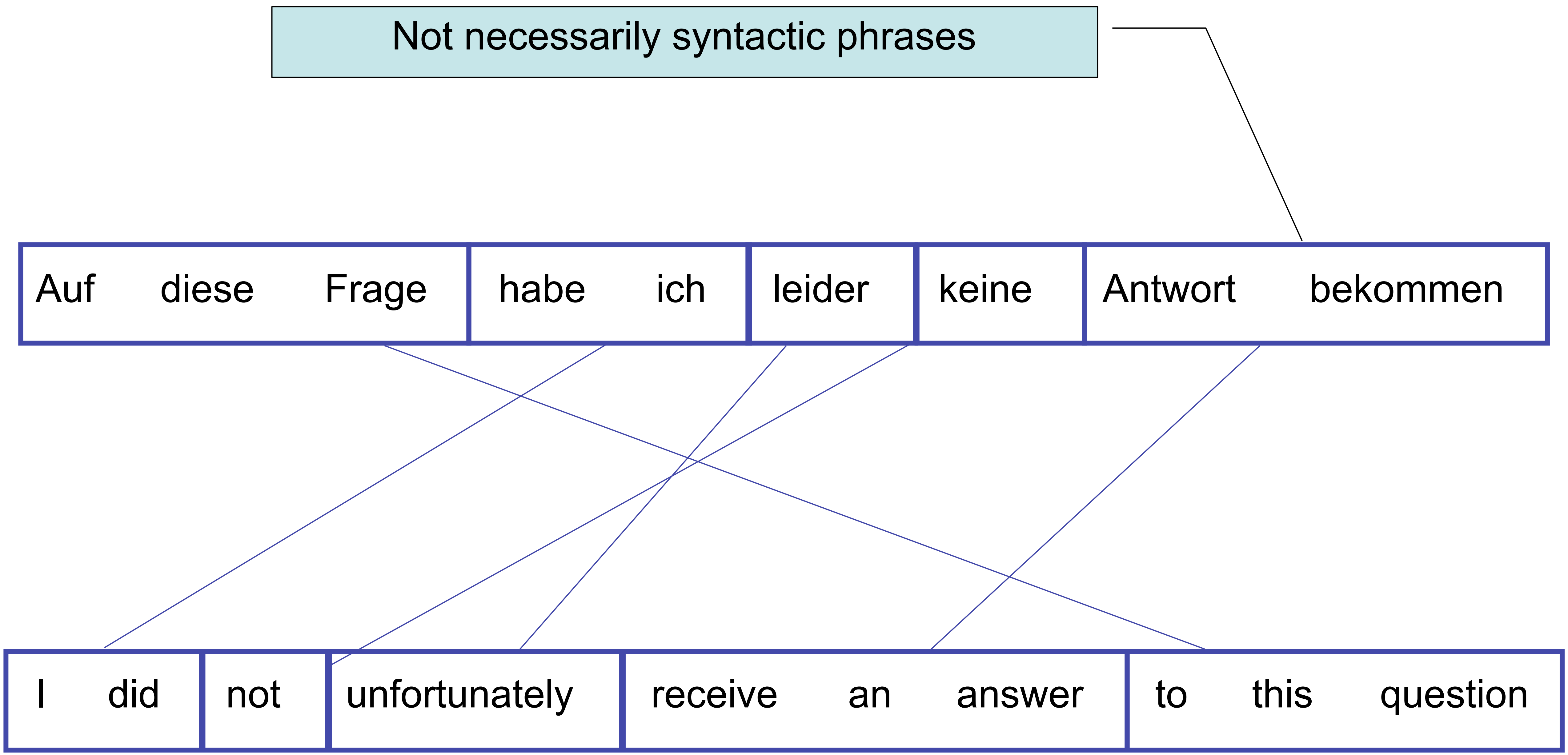
- Usually directed: each word in the target generated by one word in the source
- Many-many and null-many links allowed
- Classic IBM models of Brown et al.
- Used now mostly for word alignment, not translation



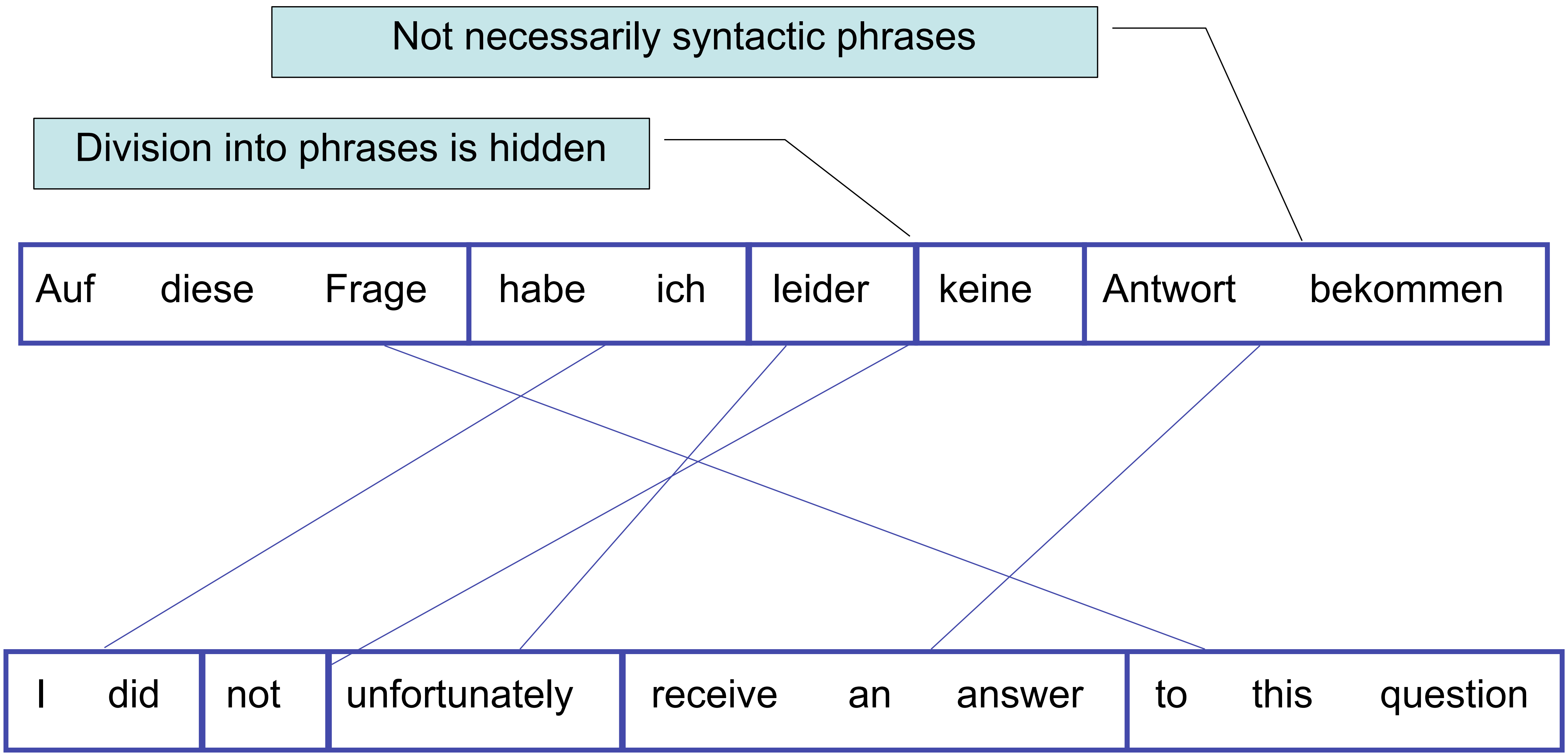
Phrase Translation Models



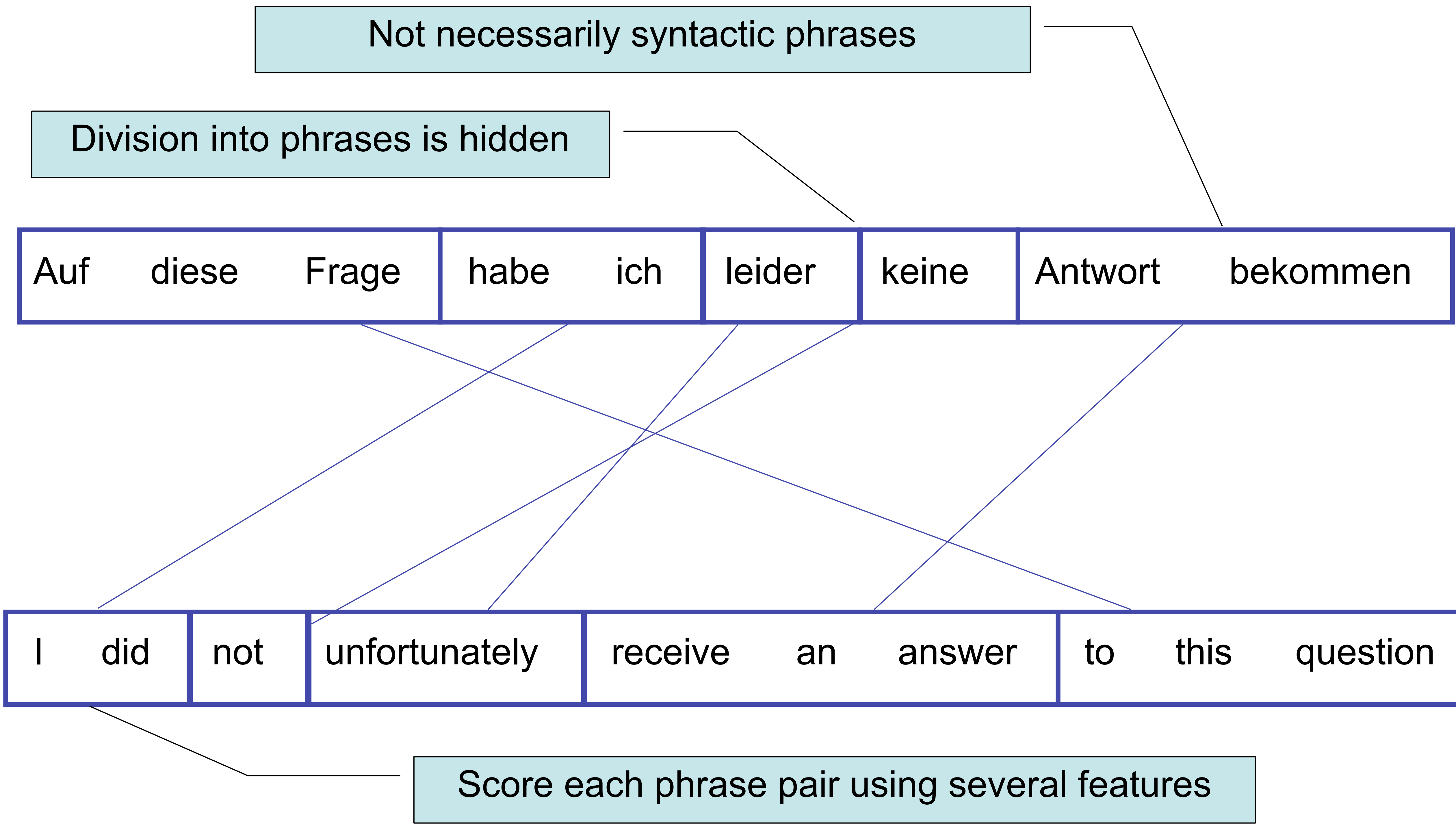
Phrase Translation Models



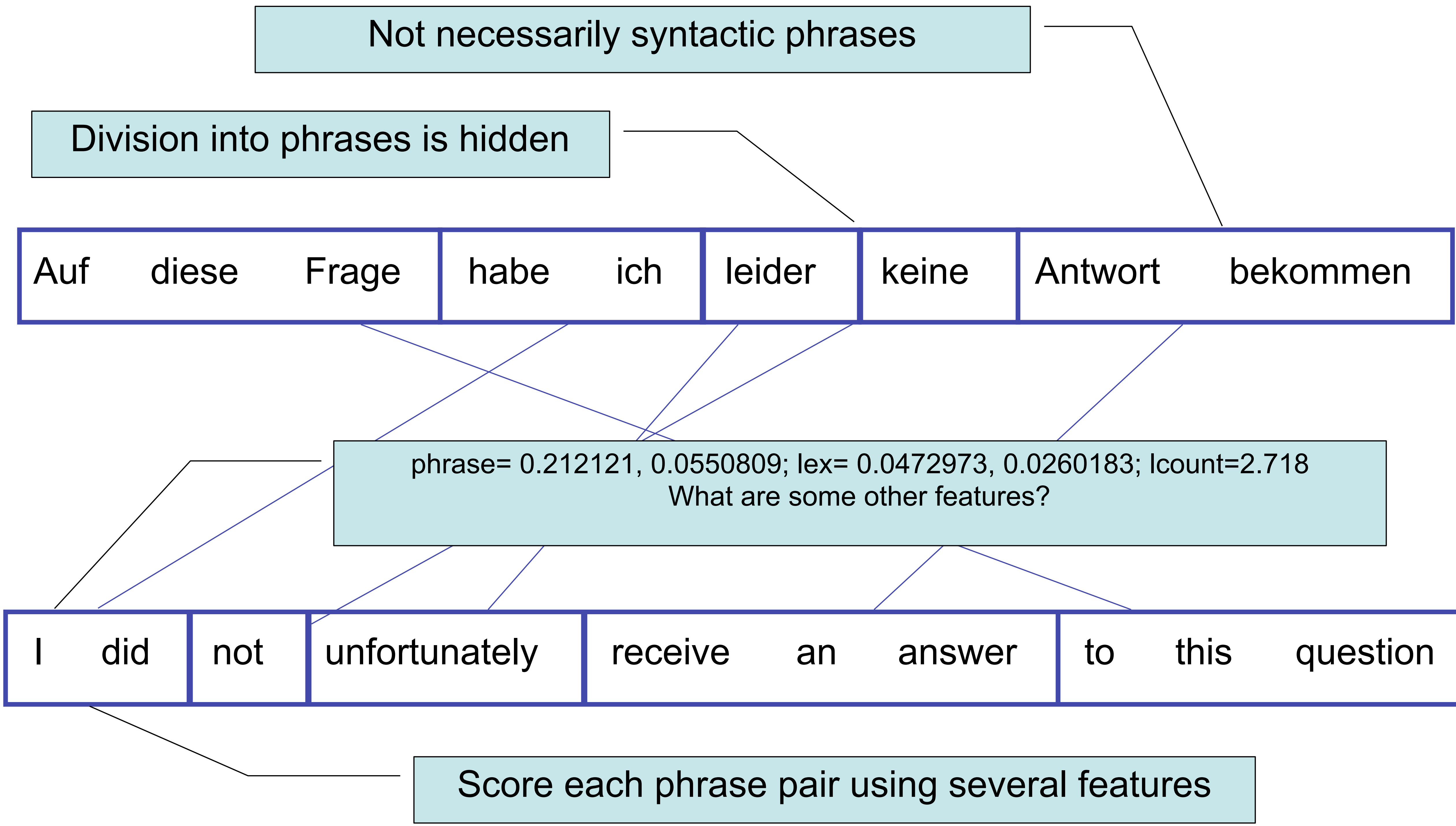
Phrase Translation Models



Phrase Translation Models



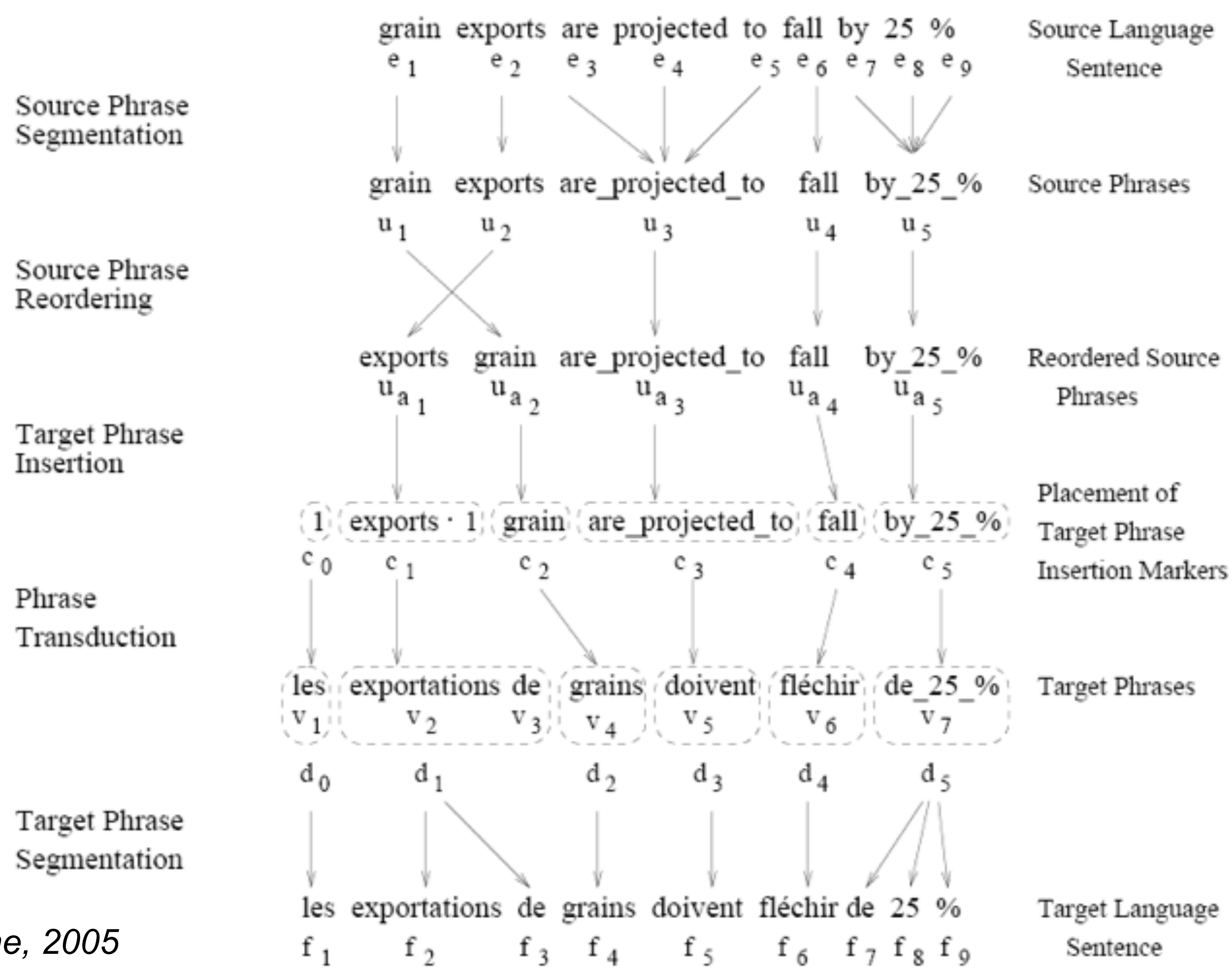
Phrase Translation Models



Phrase Translation Models

- Capture translations **in context**
  - *en Amerique*: **to** America
  - *en anglais*: **in** English
- State-of-the-art for several years
- Each source/target phrase pair is scored by several weighted features.
- The weighted sum of model features is the whole translation's score.
- Phrases don't overlap (cf. language models) but have "reordering" features.

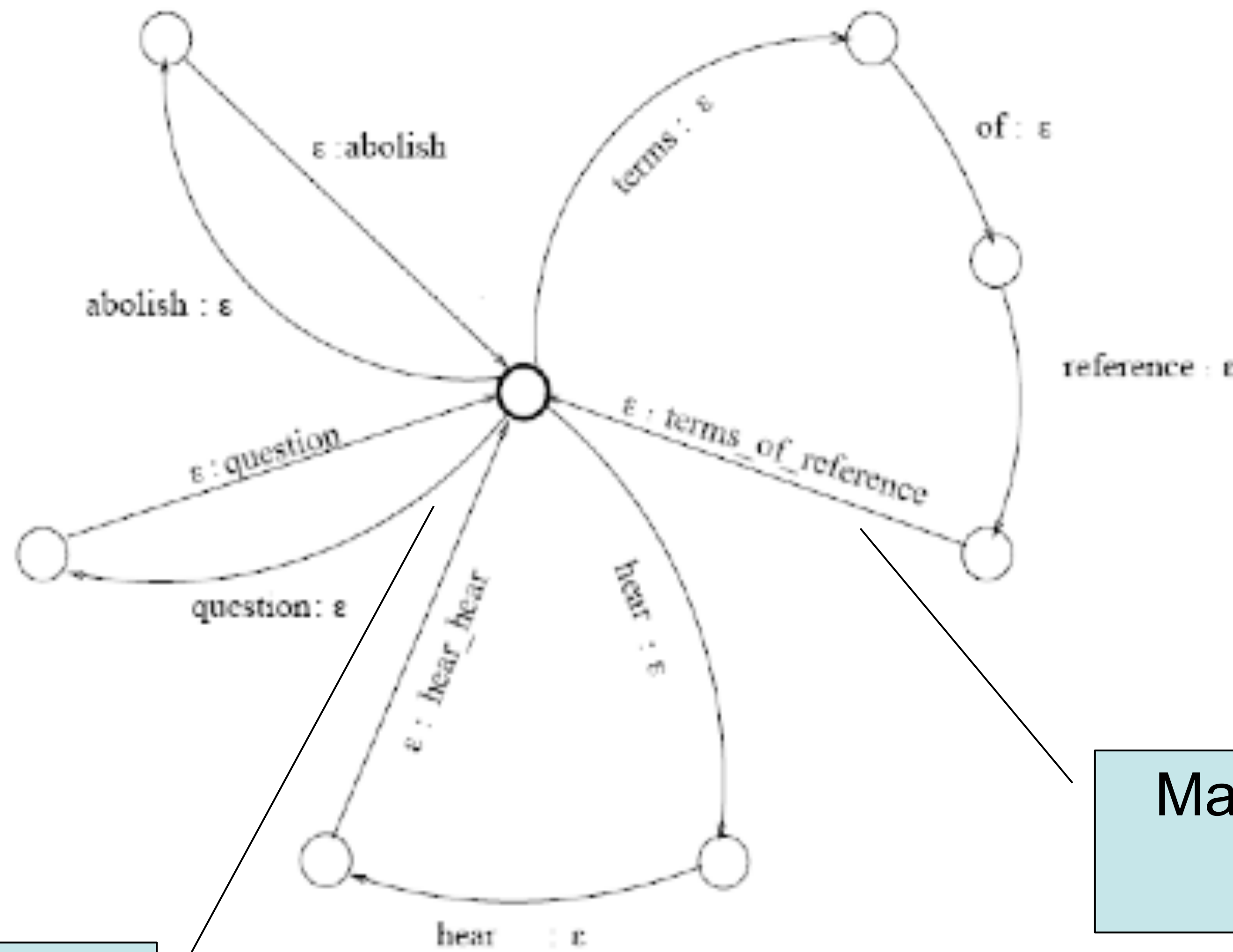
Finite State Models



Kumar, Deng & Byrne, 2005

Finite State Models

First transducer in the pipeline



Map distinct words to phrases

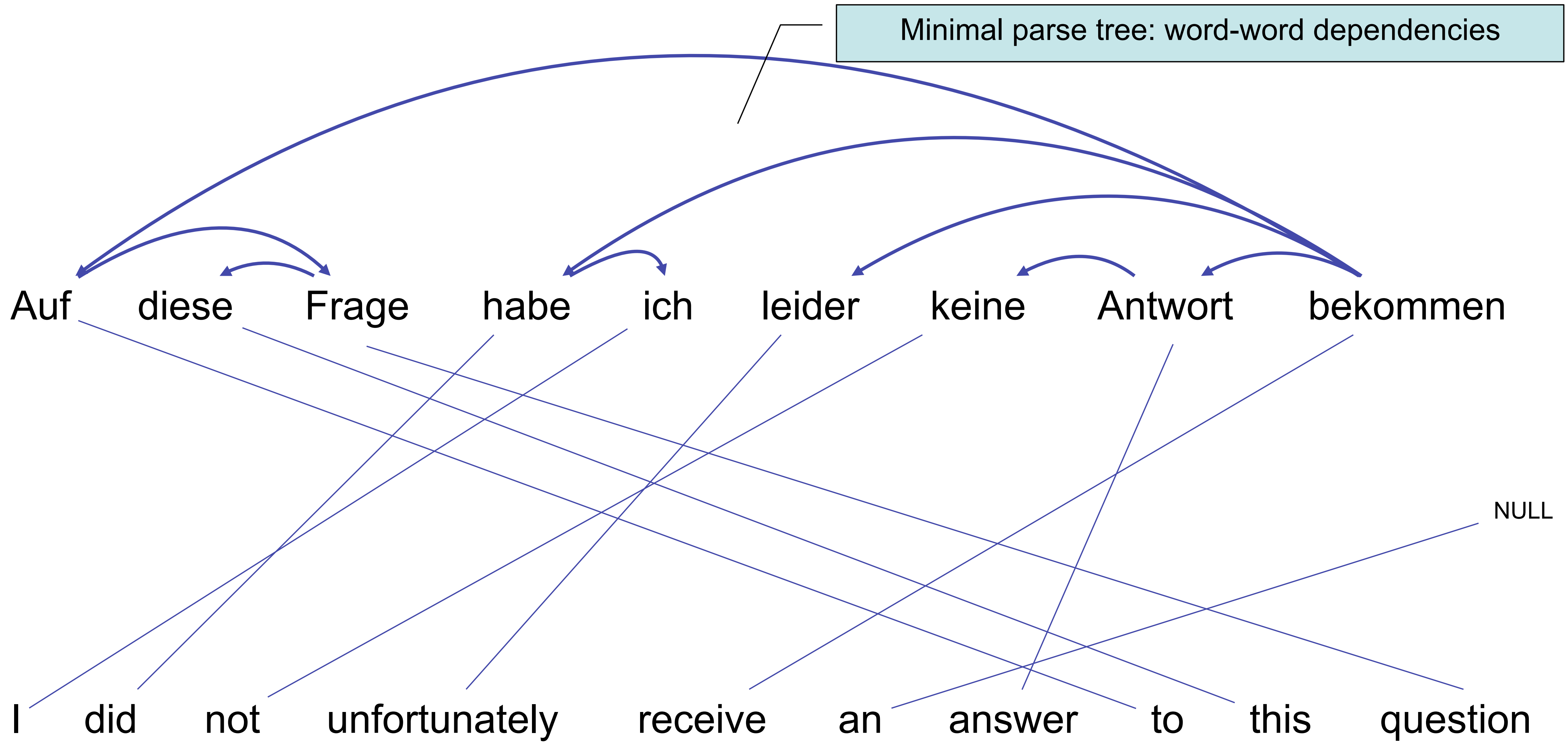
Here a unigram model of phrases

Kumar, Deng & Byrne, 2005

Finite State Models

- Natural composition with other finite state processes, e.g. Chinese word segmentation
- Standard algorithms and widely available tools (e.g. AT&T fsm toolkit)
- Limit reordering to finite offset
- Often impractical to compose all finite state machines offline

Single-Tree Translation Models

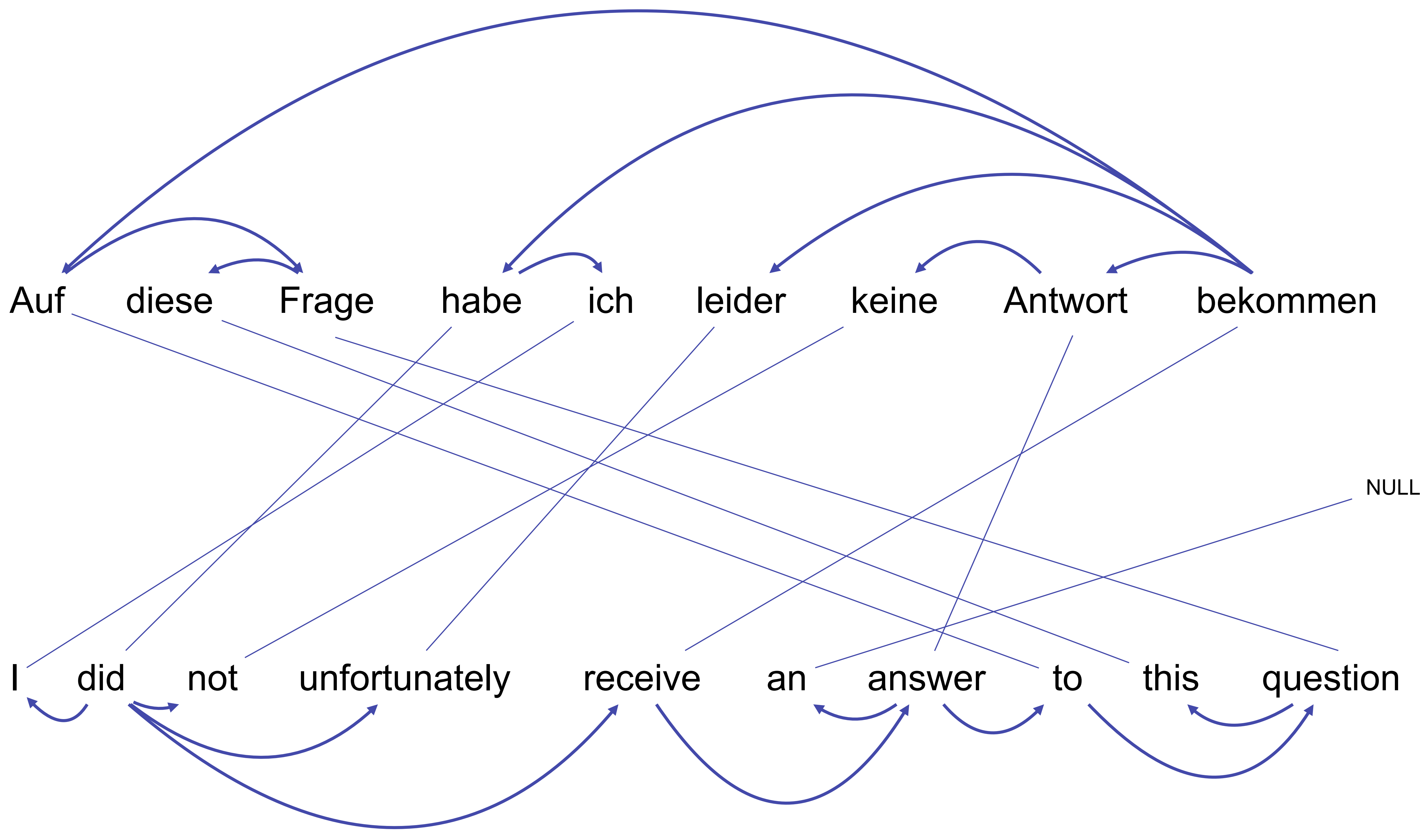


Parse trees with deeper structure have also been used.

Single-Tree Translation Models

- Either source or target has a hidden tree/parse structure
  - Also known as “tree-to-string” or “tree-transducer” models
- The side with the tree generates words/phrases in tree, not string, order.
- Nodes in the tree also generate words/phrases on the other side.
- English side is often parsed, whether it’s source or target, since English parsing is more advanced.

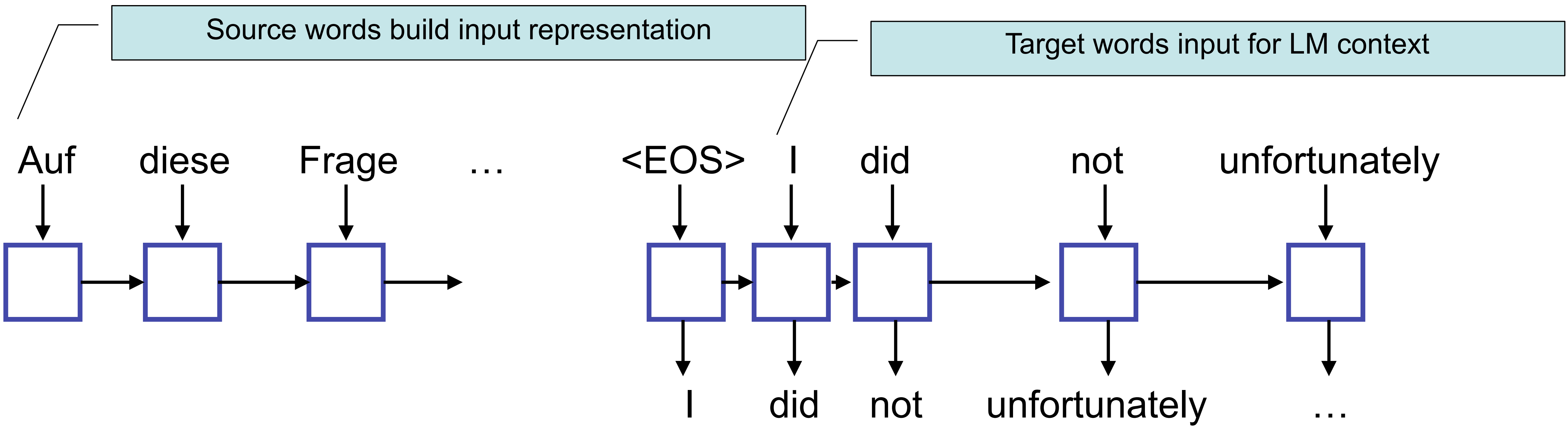
Tree-Tree Translation Models



Tree-Tree Translation Models

- Both sides have hidden tree structure
  - Can be represented with a “synchronous” grammar
- Some models assume isomorphic trees, where parent-child relations are preserved; others do not.
- Trees can be fixed in advance by monolingual parsers or induced from data (e.g. Hiero).
- Cheap trees: project from one side to the other

Latent Seq-Seq Models



- Various methods for building source representation
  - Recurrent NN, LSTM, ConvNN, Neural attention
- Representation replicated at each output position
  - Integrated LM, or combined in beam search

Learning Word Translations from Parallel Text

The “IBM Models”

Lexical translation

- How to translate a word $\rightarrow$ look up in dictionary

  Haus — *house, building, home, household, shell.*

- *Multiple translations*
  - some more frequent than others
  - for instance: *house*, and *building* most common
  - special cases: *Haus* of a *snail* is its *shell*

- Note: During all the lectures, we will translate from a foreign language into English

Collect statistics

- Look at a *parallel corpus* (German text along with English translation)

Translation of *Haus*	**Count**
house	8,000
building	1,600
home	200
household	150
shell	50

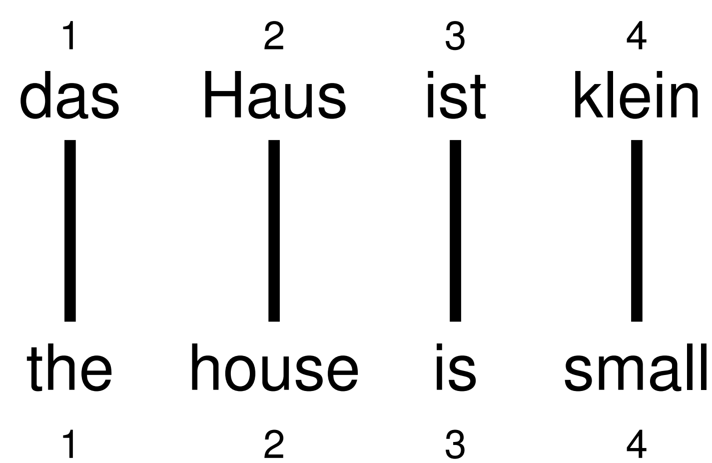
Estimate translation probabilities

- *Maximum likelihood estimation*

$$p_f(e) = \begin{cases} 0.8 & \text{if } e = \textit{house}, \\ 0.16 & \text{if } e = \textit{building}, \\ 0.02 & \text{if } e = \textit{home}, \\ 0.015 & \text{if } e = \textit{household}, \\ 0.005 & \text{if } e = \textit{shell}. \end{cases}$$

Alignment

- In a parallel text (or when we translate), we **align** words in one language with the words in the other



- Word *positions* are numbered 1–4

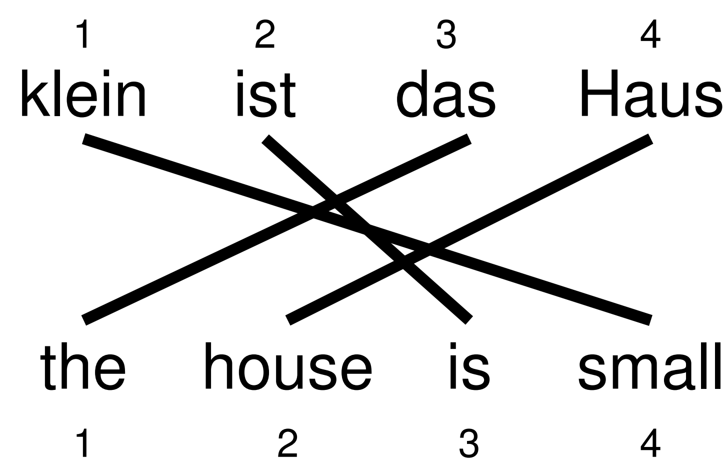
Alignment function

- Formalizing *alignment* with an **alignment function**
- Mapping an English target word at position i to a German source word at position j with a function $a : i \rightarrow j$
- Example

$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4\}$$

Reordering

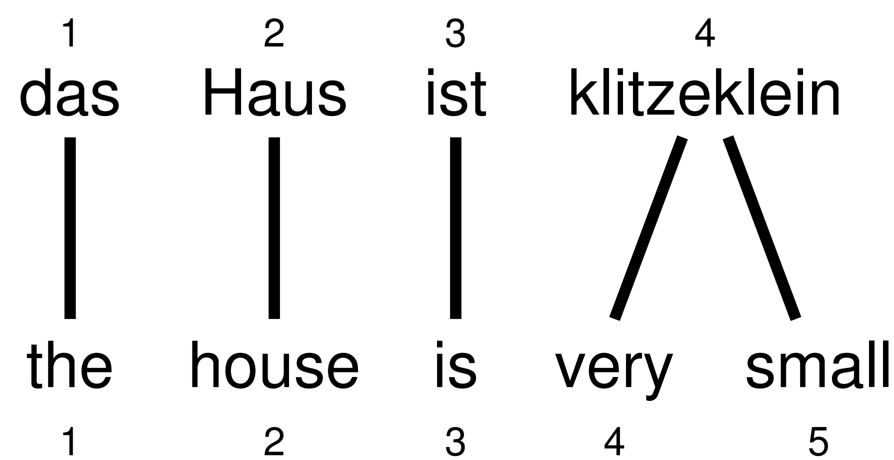
- Words may be **reordered** during translation



$$a : \{1 \rightarrow 3, 2 \rightarrow 4, 3 \rightarrow 2, 4 \rightarrow 1\}$$

One-to-many translation

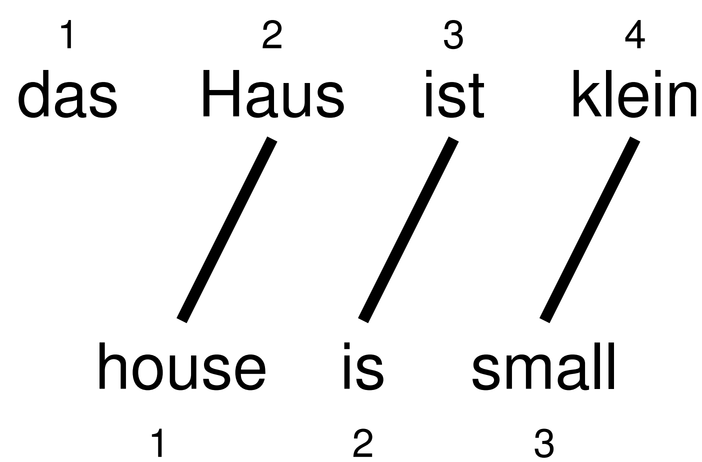
- A source word may translate into **multiple** target words



$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4, 5 \rightarrow 4\}$$

Dropping words

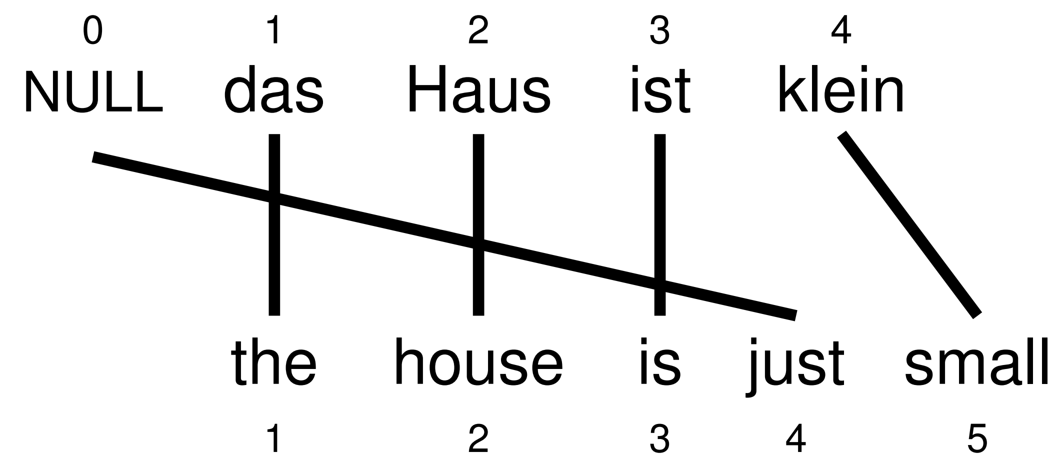
- Words may be **dropped** when translated
  - The German article *das* is dropped



$$a : \{1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 4\}$$

Inserting words

- Words may be **added** during translation
  - The English *just* does not have an equivalent in German
  - We still need to map it to something: special NULL token



$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 0, 5 \rightarrow 4\}$$

IBM Model 1

- *Generative model*: break up translation process into smaller steps
  - **IBM Model 1** only uses *lexical translation*

- Translation probability
  - for a foreign sentence $\mathbf{f} = (f_1, ..., f_{l_f})$ of length l_f
  - to an English sentence $\mathbf{e} = (e_1, ..., e_{l_e})$ of length l_e
  - with an alignment of each English word e_j to a foreign word f_i according to the alignment function $a : j \rightarrow i$

$$p(\mathbf{e}, a|\mathbf{f}) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

  - parameter ϵ is a *normalization constant*

Example

das

e	$t(e\|f)$
the	0.7
that	0.15
which	0.075
who	0.05
this	0.025

Haus

e	$t(e\|f)$
house	0.8
building	0.16
home	0.02
household	0.015
shell	0.005

ist

e	$t(e\|f)$
is	0.8
's	0.16
exists	0.02
has	0.015
are	0.005

klein

e	$t(e\|f)$
small	0.4
little	0.4
short	0.1
minor	0.06
petty	0.04

$$\begin{aligned} p(e,a|f) &= \frac{\epsilon}{4^3} \times t(\text{the}|\text{das}) \times t(\text{house}|\text{Haus}) \times t(\text{is}|\text{ist}) \times t(\text{small}|\text{klein}) \\ &= \frac{\epsilon}{4^3} \times 0.7 \times 0.8 \times 0.8 \times 0.4 \\ &= 0.0028\epsilon \end{aligned}$$

Learning lexical translation models

- We would like to *estimate* the lexical translation probabilities $t(e|f)$ from a parallel corpus

- ... but we do not have the alignments

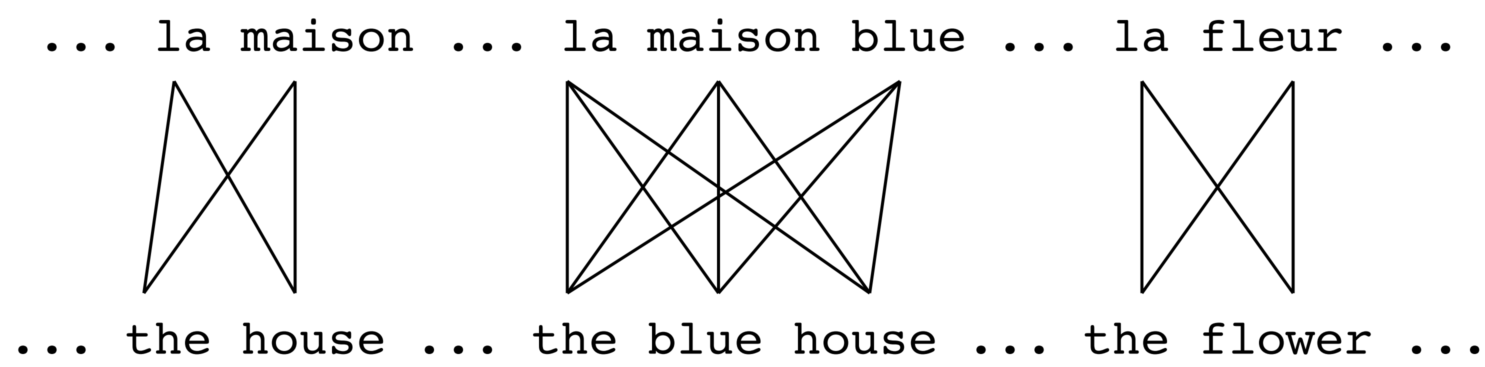
- **Chicken and egg problem**
  - if we had the *alignments*,
    $\longrightarrow$ we could estimate the *parameters* of our generative model
  - if we had the *parameters*,
    $\longrightarrow$ we could estimate the *alignments*

EM algorithm

- **Incomplete data**
  - if we had *complete data*, would could estimate *model*
  - if we had *model*, we could fill in the *gaps in the data*

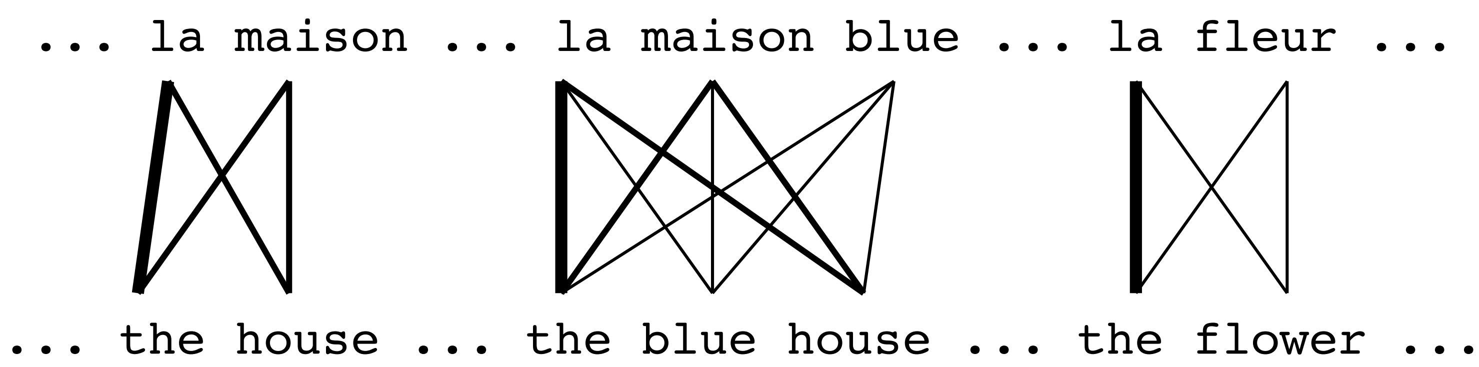
- **Expectation Maximization (EM)** in a nutshell
  - initialize model parameters (e.g. uniform)
  - assign probabilities to the missing data
  - estimate model parameters from completed data
  - iterate

EM algorithm



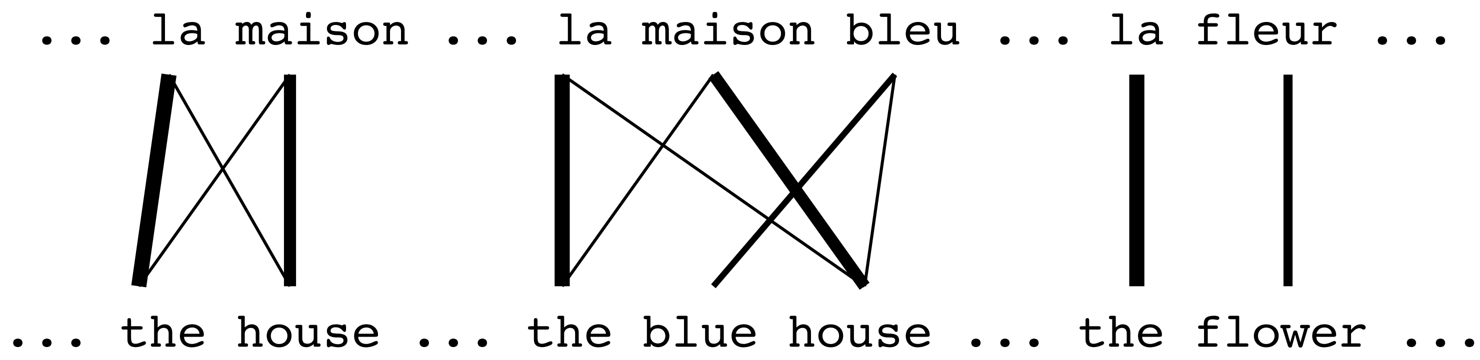
- Initial step: all alignments equally likely
- Model learns that, e.g., *la* is often aligned with *the*

EM algorithm



- After one iteration
- Alignments, e.g., between *la* and *the* are more likely

EM algorithm

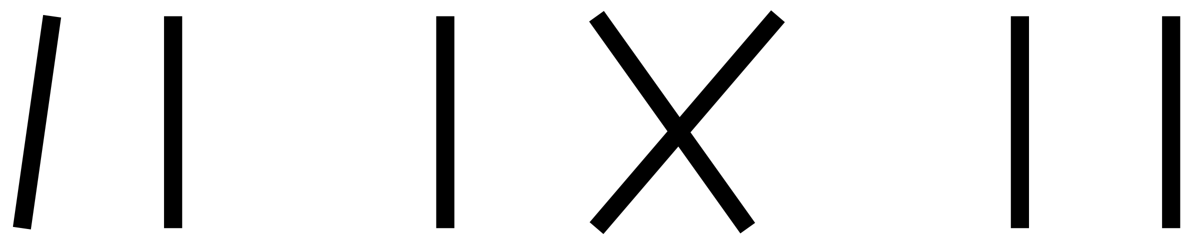


- After another iteration

- It becomes apparent that alignments, e.g., between *fleur* and *flower* are more likely (**pigeon hole principle**)

EM algorithm

`... la maison ... la maison bleu ... la fleur ...`

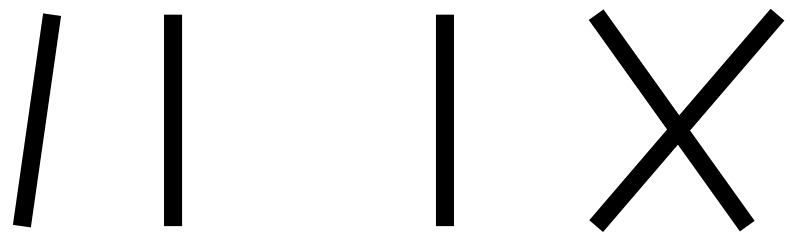


`... the house ... the blue house ... the flower ...`

- Convergence
- Inherent hidden structure revealed by EM

EM algorithm

... la maison ... la maison bleu ... la fleur ...



... the house ... the blue house ... the flower ...

p(la|the) = 0.453
p(le|the) = 0.334
p(maison|house) = 0.876
p(bleu|blue) = 0.563
...

- Parameter estimation from the aligned corpus

IBM Model 1 and EM

- EM Algorithm consists of two steps
- **Expectation-Step**: Apply model to the data
  - parts of the model are hidden (here: alignments)
  - using the model, assign probabilities to possible values
- **Maximization-Step**: Estimate model from data
  - take assign values as fact
  - collect counts (weighted by probabilities)
  - estimate model from counts
- Iterate these steps until **convergence**

IBM Model 1 and EM

- We need to be able to compute:
  - Expectation-Step: probability of alignments
  - Maximization-Step: count collection

IBM Model 1 and EM

- **Probabilities**

$$p(\text{the}|\text{la}) = 0.7 \qquad p(\text{house}|\text{la}) = 0.05$$
$$p(\text{the}|\text{maison}) = 0.1 \qquad p(\text{house}|\text{maison}) = 0.8$$

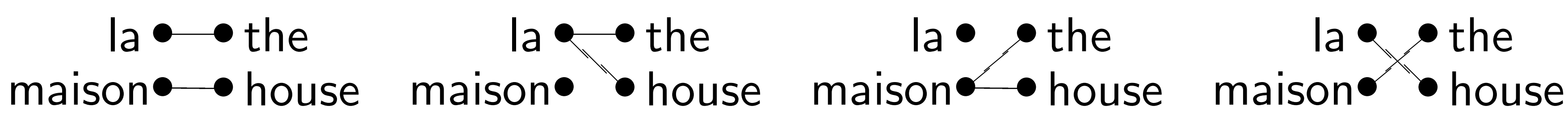
- **Alignments**

IBM Model 1 and EM

- **Probabilities**

$$p(\text{the}|\text{la}) = 0.7 \qquad p(\text{house}|\text{la}) = 0.05$$
$$p(\text{the}|\text{maison}) = 0.1 \qquad p(\text{house}|\text{maison}) = 0.8$$

- **Alignments**



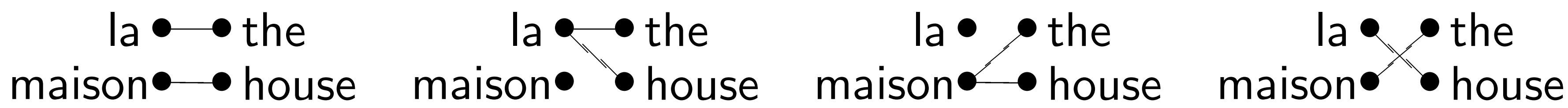
$p(\mathbf{e}, a|\mathbf{f}) = 0.56$ $\quad$ $p(\mathbf{e}, a|\mathbf{f}) = 0.035$ $\quad$ $p(\mathbf{e}, a|\mathbf{f}) = 0.08$ $\quad$ $p(\mathbf{e}, a|\mathbf{f}) = 0.005$

IBM Model 1 and EM

- **Probabilities**

$$p(\text{the}|\text{la}) = 0.7 \qquad p(\text{house}|\text{la}) = 0.05$$
$$p(\text{the}|\text{maison}) = 0.1 \qquad p(\text{house}|\text{maison}) = 0.8$$

- **Alignments**



$$p(\mathbf{e}, a|\mathbf{f}) = 0.56 \qquad p(\mathbf{e}, a|\mathbf{f}) = 0.035 \qquad p(\mathbf{e}, a|\mathbf{f}) = 0.08 \qquad p(\mathbf{e}, a|\mathbf{f}) = 0.005$$

$$p(a|\mathbf{e}, \mathbf{f}) = 0.824 \qquad p(a|\mathbf{e}, \mathbf{f}) = 0.052 \qquad p(a|\mathbf{e}, \mathbf{f}) = 0.118 \qquad p(a|\mathbf{e}, \mathbf{f}) = 0.007$$

- **Counts**

$$c(\text{the}|\text{la}) = 0.824 + 0.052 \qquad c(\text{house}|\text{la}) = 0.052 + 0.007$$
$$c(\text{the}|\text{maison}) = 0.118 + 0.007 \qquad c(\text{house}|\text{maison}) = 0.824 + 0.118$$

IBM Model 1 and EM: Expectation Step

- We need to compute $p(a|\mathbf{e},\mathbf{f})$
- Applying the *chain rule*:

$$p(a|\mathbf{e},\mathbf{f}) = \frac{p(\mathbf{e},a|\mathbf{f})}{p(\mathbf{e}|\mathbf{f})}$$

- We already have the formula for $p(\mathbf{e},\mathbf{a}|\mathbf{f})$ (definition of Model 1)

IBM Model 1 and EM: Expectation Step

- We need to compute $p(\mathbf{e}|\mathbf{f})$

$$p(\mathbf{e}|\mathbf{f}) = \sum_a p(\mathbf{e}, a|\mathbf{f})$$
$$= \sum_{a(1)=0}^{l_f} \dots \sum_{a(l_e)=0}^{l_f} p(\mathbf{e}, a|\mathbf{f})$$
$$= \sum_{a(1)=0}^{l_f} \dots \sum_{a(l_e)=0}^{l_f} \frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

IBM Model 1 and EM: Expectation Step

$$p(\mathbf{e}|\mathbf{f}) = \sum_{a(1)=0}^{l_f} ... \sum_{a(l_e)=0}^{l_f} \frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

$$= \frac{\epsilon}{(l_f+1)^{l_e}} \sum_{a(1)=0}^{l_f} ... \sum_{a(l_e)=0}^{l_f} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

$$= \frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_j|f_i)$$

- Note the trick in the last line
  - removes the need for an *exponential* number of products

$\rightarrow$ this makes IBM Model 1 estimation **tractable**

IBM Model 1 and EM: Expectation Step

- Combine what we have:

$$\begin{aligned} p(\mathbf{a}|\mathbf{e},\mathbf{f}) &= p(\mathbf{e},\mathbf{a}|\mathbf{f})/p(\mathbf{e}|\mathbf{f}) \\ &= \frac{\frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})}{\frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_j|f_i)} \\ &= \prod_{j=1}^{l_e} \frac{t(e_j|f_{a(j)})}{\sum_{i=0}^{l_f} t(e_j|f_i)} \end{aligned}$$

IBM Model 1 and EM: Maximization Step

- Now we have to *collect counts*

- Evidence from a sentence pair **e,f** that word e is a translation of word f:

$$c(e|f;\mathbf{e},\mathbf{f}) = \sum_a p(a|\mathbf{e},\mathbf{f}) \sum_{j=1}^{l_e} \delta(e,e_j)\delta(f,f_{a(j)})$$

- With the same simplication as before:

$$c(e|f;\mathbf{e},\mathbf{f}) = \frac{t(e|f)}{\sum_{j=1}^{l_e} t(e|f_{a(j)})} \sum_{j=1}^{l_e} \delta(e,e_j) \sum_{i=0}^{l_f} \delta(f,f_i)$$

IBM Model 1 and EM: Maximization Step

- After collecting these counts over a corpus, we can estimate the model:

$$t(e|f;\mathbf{e},\mathbf{f}) = \frac{\sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f}))}{\sum_f \sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f}))}$$

IBM Model 1 and EM: Pseudocode

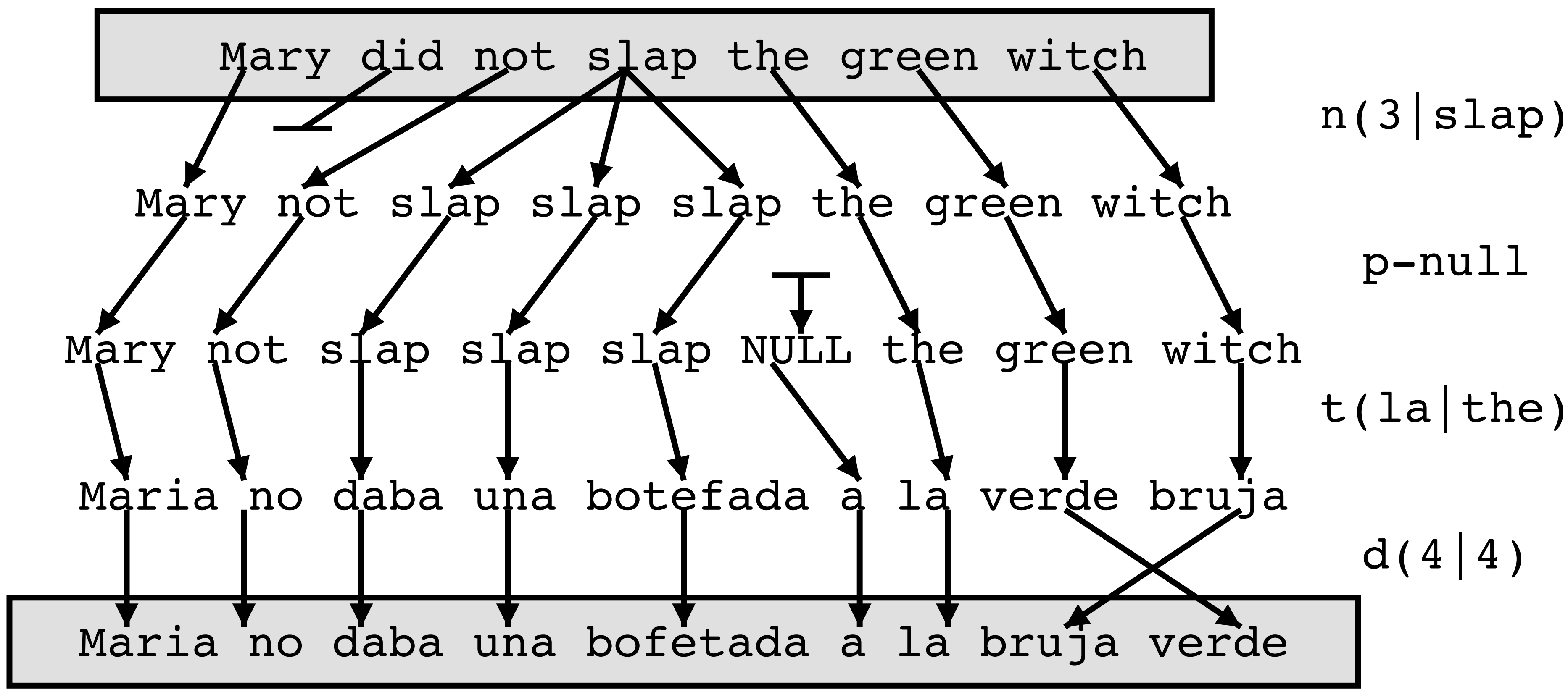
```
initialize t(e|f) uniformly
do
  set count(e|f) to 0 for all e,f
  set total(f) to 0 for all f
  for all sentence pairs (e_s,f_s)
    for all words e in e_s
      total_s = 0
      for all words f in f_s
        total_s += t(e|f)
    for all words e in e_s
      for all words f in f_s
        count(e|f) += t(e|f) / total_s
        total(f)   += t(e|f) / total_s
  for all f in domain( total(.) )
    for all e in domain( count(.|f) )
      t(e|f) = count(e|f) / total(f)
until convergence
```

Higher IBM Models

IBM Model 1	lexical translation
IBM Model 2	adds absolute **reordering model**
IBM Model 3	adds **fertility model**
IBM Model 4	relative reordering model
IBM Model 5	fixes **deficiency**

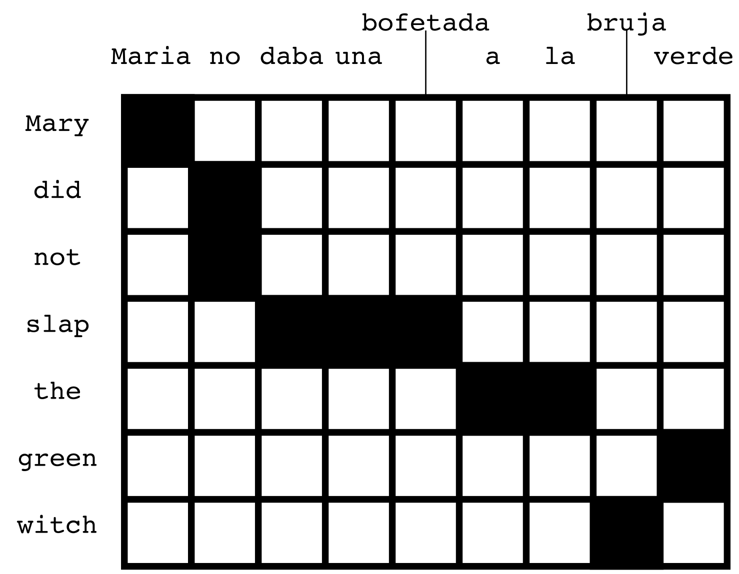
- Only IBM Model 1 has *global maximum*
  - training of a higher IBM model builds on previous model
- Compuationally biggest change in Model 3
  - trick to simplify estimation does not work anymore
  - → *exhaustive* count collection becomes computationally too expensive
  - **sampling** over high probability alignments is used instead

IBM Model 4


Mary did not slap the green witch
n(3|slap)
Mary not slap slap slap the green witch
p-null
Mary not slap slap slap NULL the green witch
t(la|the)
Maria no daba una botefada a la verde bruja
d(4|4)
Maria no daba una bofetada a la bruja verde

Word alignment

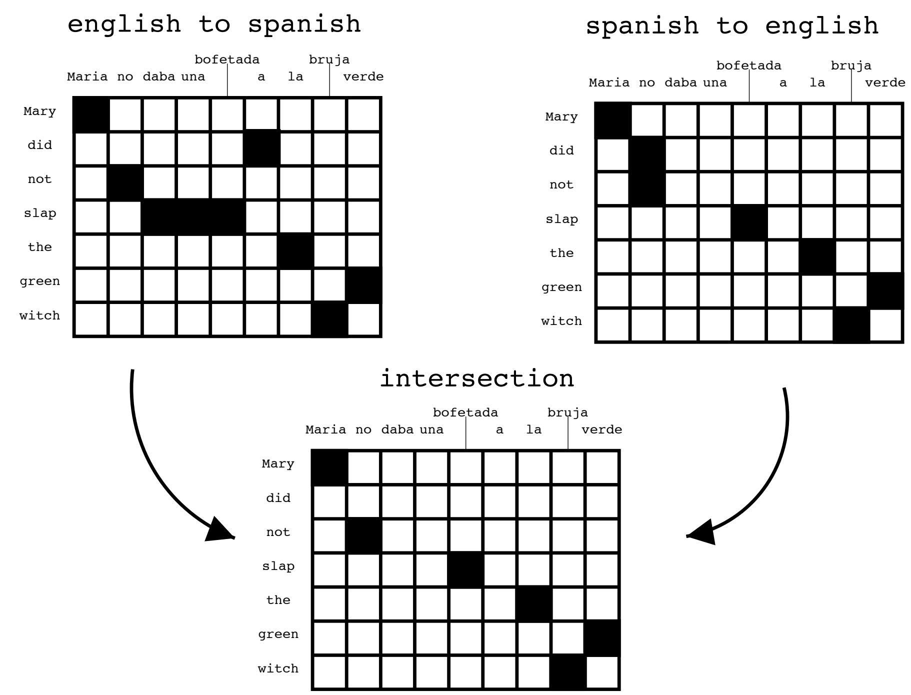
- Notion of **word alignment** valuable
- Shared task at NAACL 2003 and ACL 2005 workshops



Word alignment with IBM models

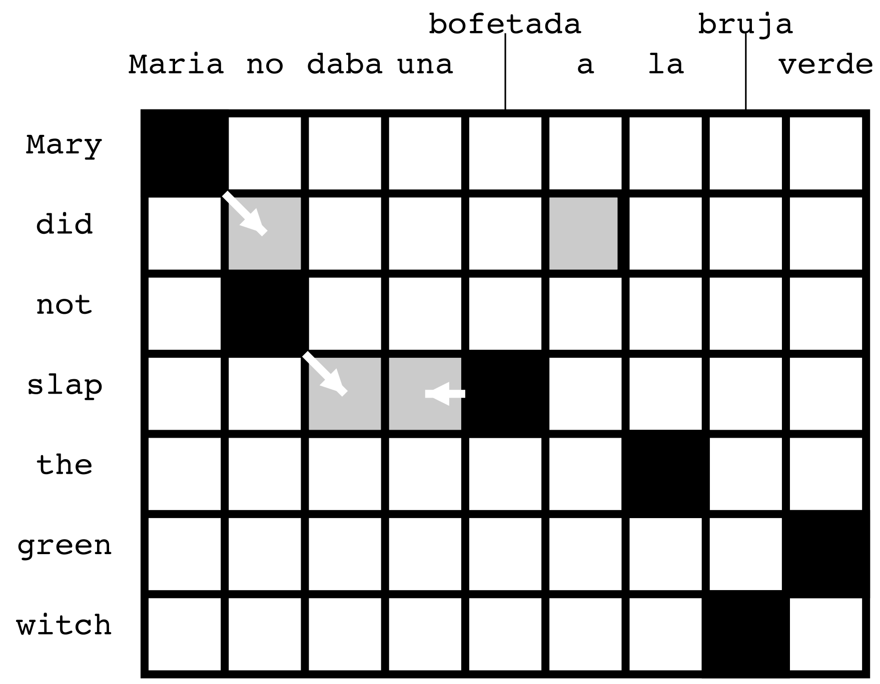
- IBM Models create a *many-to-one* mapping
  - words are aligned using an **alignment function**
  - a function may return the same value for different input (one-to-many mapping)
  - a function can not return multiple values for one input (*no many-to-one* mapping)
- But we need *many-to-many* mappings

Symmetrizing word alignments



- *Intersection* of GIZA++ bidirectional alignments

Symmetrizing word alignments



- *Grow* additional alignment points [Och and Ney, CompLing2003]

Growing heuristic

```
GROW-DIAG-FINAL(e2f,f2e):
  neighboring = ((-1,0),(0,-1),(1,0),(0,1),(-1,-1),(-1,1),(1,-1),(1,1))
  alignment = intersect(e2f,f2e);
  GROW-DIAG(); FINAL(e2f); FINAL(f2e);

GROW-DIAG():
  iterate until no new points added
    for english word e = 0 ... en
      for foreign word f = 0 ... fn
        if ( e aligned with f )
          for each neighboring point ( e-new, f-new ):
            if ( ( e-new not aligned and f-new not aligned ) and
                 ( e-new, f-new ) in union( e2f, f2e ) )
              add alignment point ( e-new, f-new )
FINAL(a):
  for english word e-new = 0 ... en
    for foreign word f-new = 0 ... fn
      if ( ( e-new not aligned or f-new not aligned ) and
           ( e-new, f-new ) in alignment a )
        add alignment point ( e-new, f-new )
```

Synchronous Grammars: Inversion Transduction Grammar

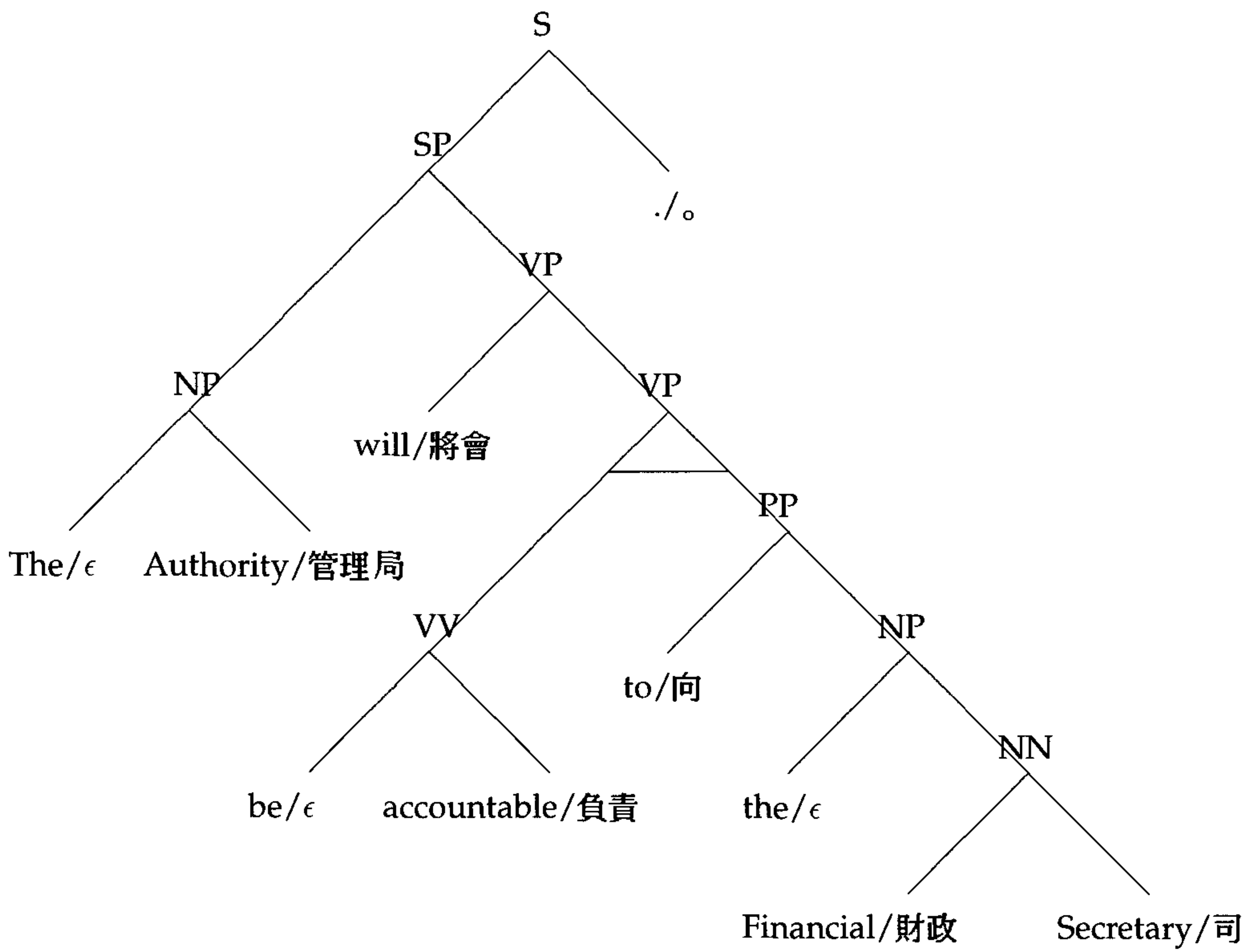
Syntactically-Motivated Distortion

The Authority will be accountable to the Financial Secretary.

管理局將會向財政司負責。

(*Authority will to Financial Secretary accountable.*)

Syntactically-Motivated Distortion



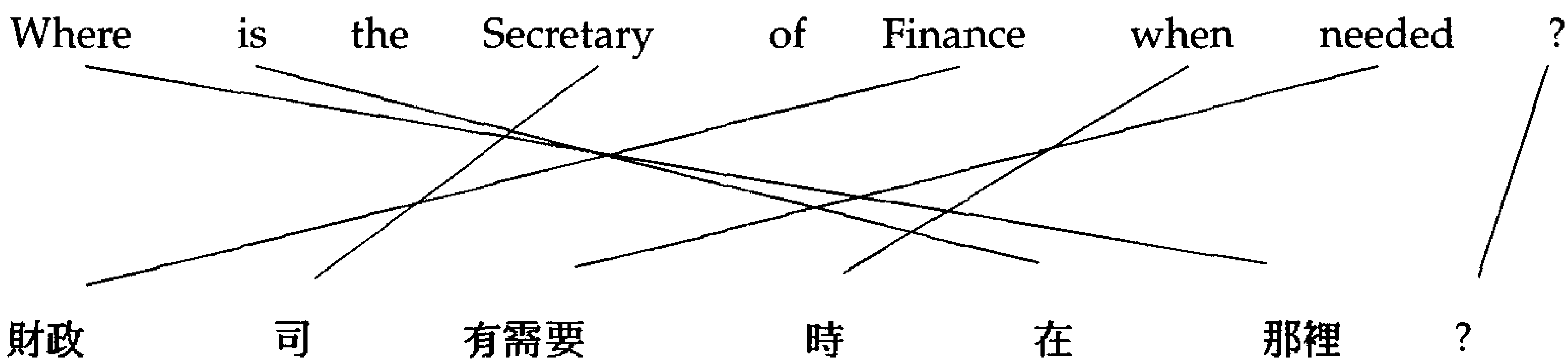
ITG Overview

- Special case of synchronous CFG
- One, joint nonterminal per bilingual node
- Children are translated monotonically, or reversed
- Binarized normal form
- Mostly used for exact, polytime alignment

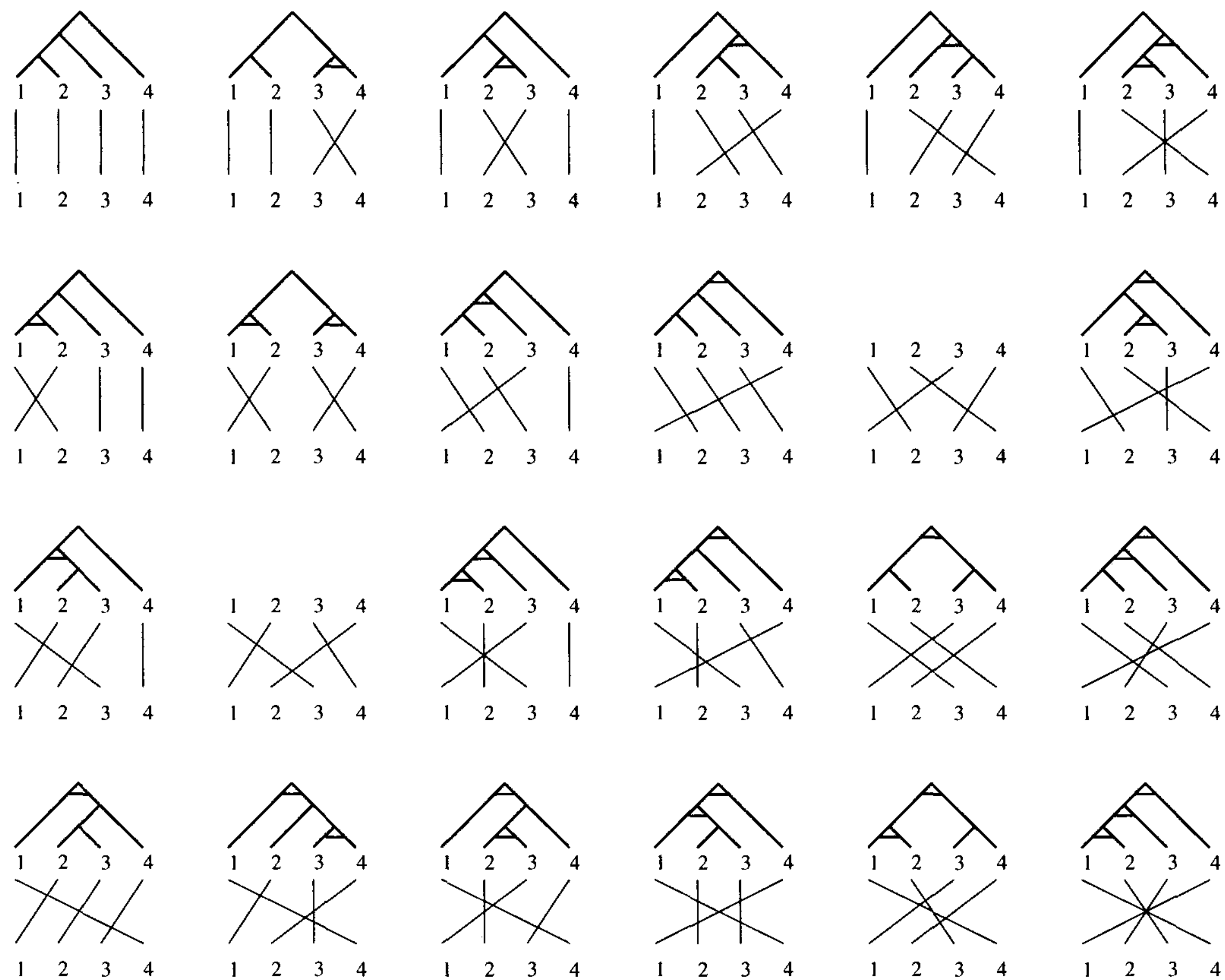
ITG Rules

S	$\rightarrow$	[SP Stop]
SP	$\rightarrow$	[NP VP] | [NP VV] | [NP V]
PP	$\rightarrow$	[Prep NP]
NP	$\rightarrow$	[Det NN] | [Det N] | [Pro] | [NP Conj NP]
NN	$\rightarrow$	[A N] | [NN PP]
VP	$\rightarrow$	[Aux VP] | [Aux VV] | [VV PP]
VV	$\rightarrow$	[V NP] | [Cop A]
Det	$\rightarrow$	the/ϵ
Prep	$\rightarrow$	to/向
Pro	$\rightarrow$	I/我 | you/你
N	$\rightarrow$	authority/管理局 | secretary/司
A	$\rightarrow$	accountable/負責 | financial/財政
Conj	$\rightarrow$	and/和
Aux	$\rightarrow$	will/將會
Cop	$\rightarrow$	be/ϵ
Stop	$\rightarrow$	./。
VP	$\rightarrow$	$\langle$VV PP$\rangle$

ITG Alignment



Legal ITG Alignments



Bracketing ITG

$$
\begin{array}{lcll}
\mathrm{A} & \xrightarrow{a} & [\mathrm{A}\ \mathrm{A}] & \\
\mathrm{A} & \xrightarrow{a} & \langle \mathrm{A}\ \mathrm{A} \rangle & \\
\mathrm{A} & \xrightarrow{b_{ij}} & u_i/v_j & \text{for all } i,j \text{ English-Chinese lexical translations} \\
\mathrm{A} & \xrightarrow{b_{i\epsilon}} & u_i/\epsilon & \text{for all } i \text{ English vocabulary} \\
\mathrm{A} & \xrightarrow{b_{\epsilon j}} & \epsilon/v_j & \text{for all } j \text{ Chinese vocabulary}
\end{array}
$$

Removing Spurious Ambiguity

$$
\begin{array}{lcll}
A & \xrightarrow{a} & [A\ B] & \\
A & \xrightarrow{a} & [B\ B] & \\
A & \xrightarrow{a} & [C\ B] & \\
A & \xrightarrow{a} & [A\ C] & \\
A & \xrightarrow{a} & [B\ C] & \\
B & \xrightarrow{a} & \langle A\ A \rangle & \\
B & \xrightarrow{a} & \langle B\ A \rangle & \\
B & \xrightarrow{a} & \langle C\ A \rangle & \\
B & \xrightarrow{a} & \langle A\ C \rangle & \\
B & \xrightarrow{a} & \langle B\ C \rangle & \\
C & \xrightarrow{b_{ij}} & u_i/v_j & \text{for all } i,j \text{ English-Chinese lexical translations} \\
C & \xrightarrow{b_{i\epsilon}} & u_i/\epsilon & \text{for all } i \text{ English vocabulary} \\
C & \xrightarrow{b_{\epsilon j}} & \epsilon/v_j & \text{for all } j \text{ Chinese vocabulary}
\end{array}
$$

Specialized Translation Models: Named Entities

Translating Words in a Sentence

- Models will automatically learn entries in probabilistic translation dictionaries, for instance p(elle|she), from co-occurrences in aligned sentences of a parallel text.

- For some kinds of words/phrases, this is less effective. For example:
  - numbers
  - dates
  - named entities (NE)

The reason: these constitute a large open class of words that will not all occur even in the largest bitext. Plus, there are regularities in translation of numbers/dates/ NE.

Handling Named Entities

- For many language pairs, and particularly those which do not share an alphabet, transliteration of person and place names is the desired method of translation.

- General Method:
    1. Identify NE's via classifier
    2. Transliterate name
    3. Translate/reorder honorifics

- Also useful for alignment. Consider the case of Inuktitut-English alignment, where Inuktitut renderings of European names are highly nondeterministic.

Transliteration

Inuktitut rendering of English names changes the string significantly but not deterministically

Williams	**McLean**
ailiams	makalain
uialims	makkalain
uilialums	maklaain
uiliam	maklain
uiliammas	maklainn
uiliams	maklait
uilians	makli
uliams	maklii
viliams	makliik
	makliin
Campbell	maklin
kaampu	malain
kaampul	matliin
kaamvul	miklain
kamvul	mikliin
	miklin

Transliteration

Inuktitut rendering of English names changes the string significantly but not deterministically

Train a **probabilistic finite-state transducer** to model this ambiguous transformation

Williams	**McLean**
ailiams	makalain
uialims	makkalain
uilialums	maklaain
uiliam	maklain
uiliammas	maklainn
uiliams	maklait
uilians	makli
uliams	maklii
viliams	makliik
	makliin
Campbell	maklin
kaampu	malain
kaampul	matliin
kaamvul	miklain
kamvul	mikliin
	miklin

Transliteration

Inuktitut rendering of English names changes the string significantly but not deterministically

Williams	McLean
ailiams	makalain
uialims	makkalain
uilialums	maklaain
uiliam	maklain
uiliammas	maklainn
uiliams	maklait
uilians	makli
uliams	maklii
viliams	makliik
	makliin
Campbell	maklin
kaampu	malain
kaampul	matliin
kaamvul	miklain
kamvul	mikliin
	miklin

… Mr. Williams … … mista uialims …

Useful Types of Word Analysis

- **Number/Date Handling**

- **Named Entity Tagging/Transliteration**

- **Morphological Analysis**
  - **Analyze a word to its root form (at least for word alignment)**

    was -> is **believing -> believe**

    ruminerai -> ruminer **ruminiez -> ruminer**
  - **As a dimensionality reduction technique**
  - **To allow lookup in existing dictionary**

Learning Word Translation Dictionaries Using Minimal Resources

Learning Translation Lexicons for Low-Resource Languages

{Serbian Uzbek Romanian Bengali} → English

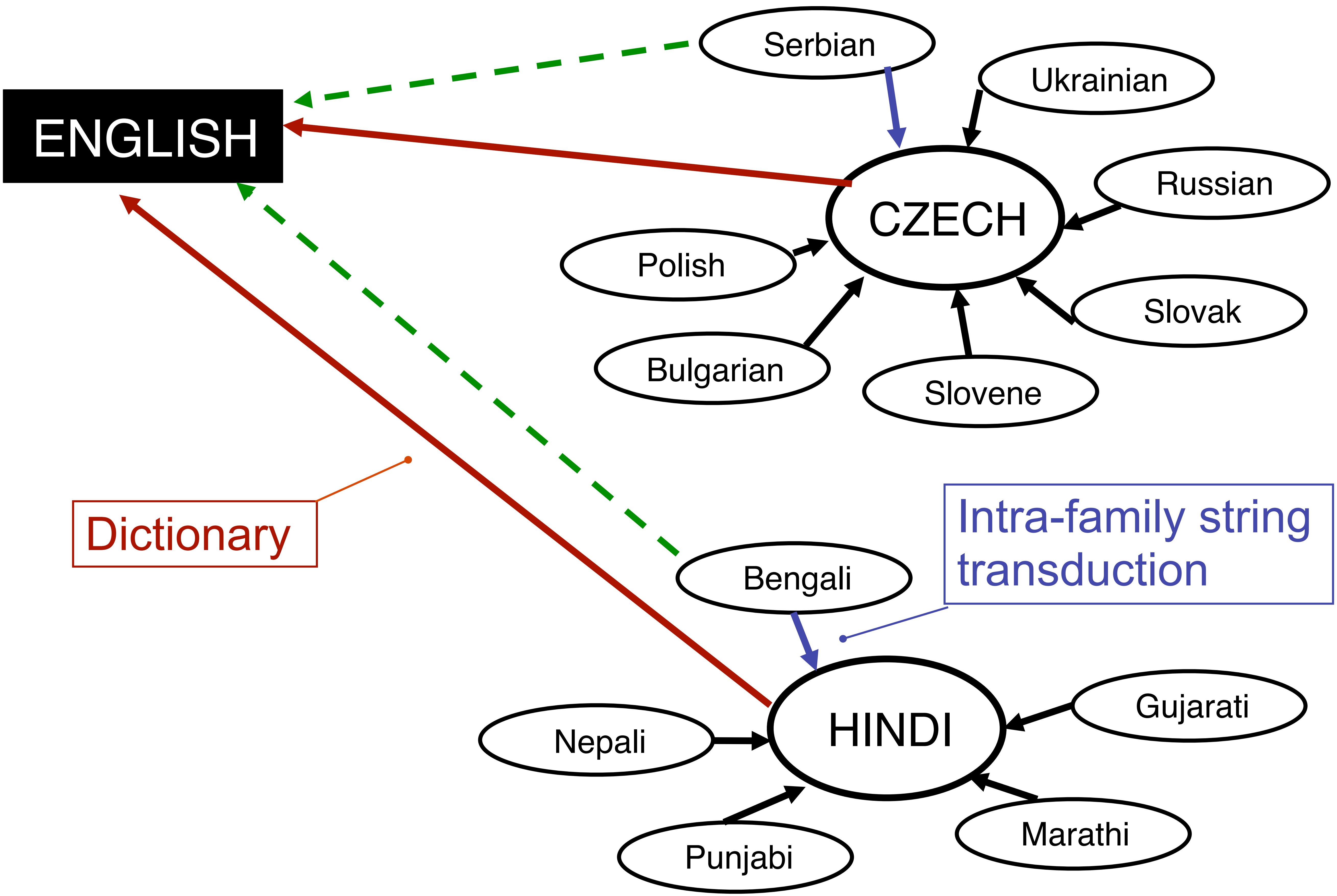
Problem: Scarce resources . . .

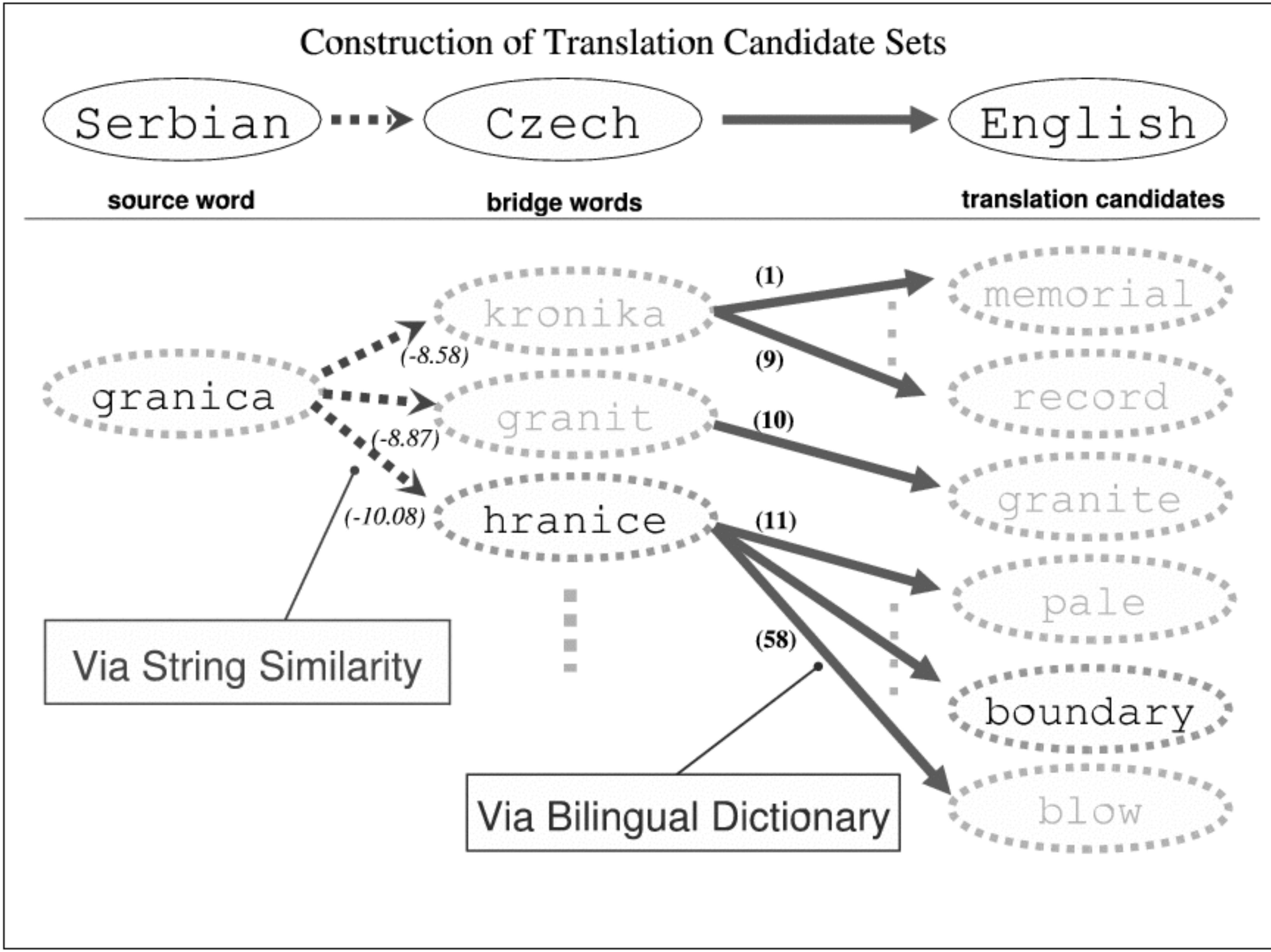
– Large parallel texts are very helpful, but often unavailable
– Often, no "seed" translation lexicon is available
– Neither are resources such as parsers, taggers, thesauri

Solution: Use only monolingual corpora in source, target languages

– But use many information sources to propose and rank translation candidates

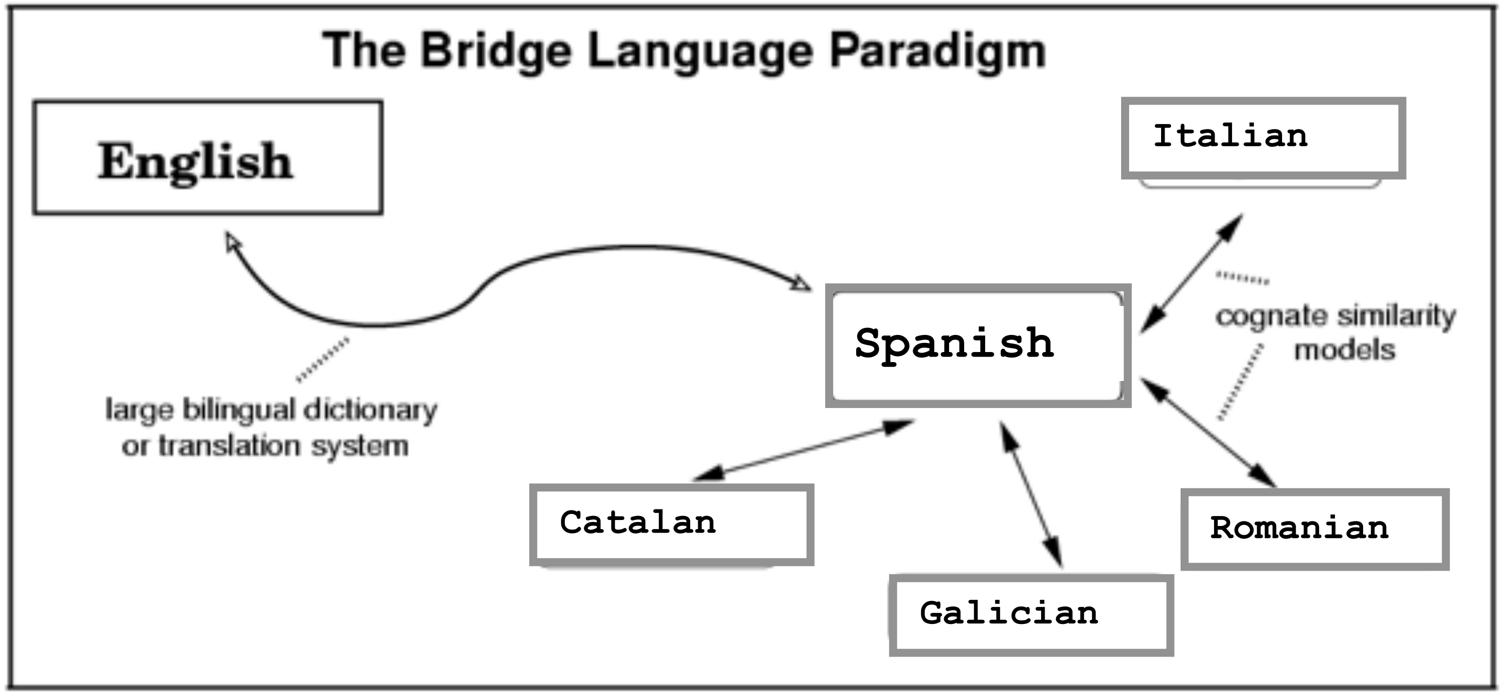
Bridge Languages





*** Constructing translation candidate sets**

Cognate Selection



some cognates

Spanish-Italian	homogenizar omogeneizzare
Polish-Serbian	befsztyk biftek
German-Dutch	gefestigt gevestigd

Spanish Word	Italian Word	Cognate?
electron	elettrone	
aventurero	avventuriero	
perífrasis	perifrasi	
divulgar	divulgare	
triada	triade	
agresivo	aggressivo	
insertar	inserto	
esprint	sprint	
trópico	tropico	
altimetro	altimetro	
alegato	lista	No
variado	variato	
cepillar	piallare	
confusin	confusione	
fortificacion	fortificazione	
conjuncion	congiunzione	
encantador	incantatore	
heredero	erede	
vidrio	vetro	
vaciar	variare	No
talisman	talismano	
sólido	solido	
criptografia	crittografia	
carencia	carenza	
cortesania	cortesia	No
sadico	sadico	
concentracion	concentrazione	
venida	venuta	
agonizante	agonizzante	
extinguir	estinguere	

The Transliteration Problem

Arabic

Piedade	BEH YEH YEH DAL ALEF DAL YEH
Bolivia	BEH WAW LAM YEH FEH YEH ALEF
Luxembourg	LAM KAF SEEN MEEM BEH WAW REH GHAIN
Zanzibar	ZAIN NOON JEEM YEH BEH ALEF REH

Inuktitut

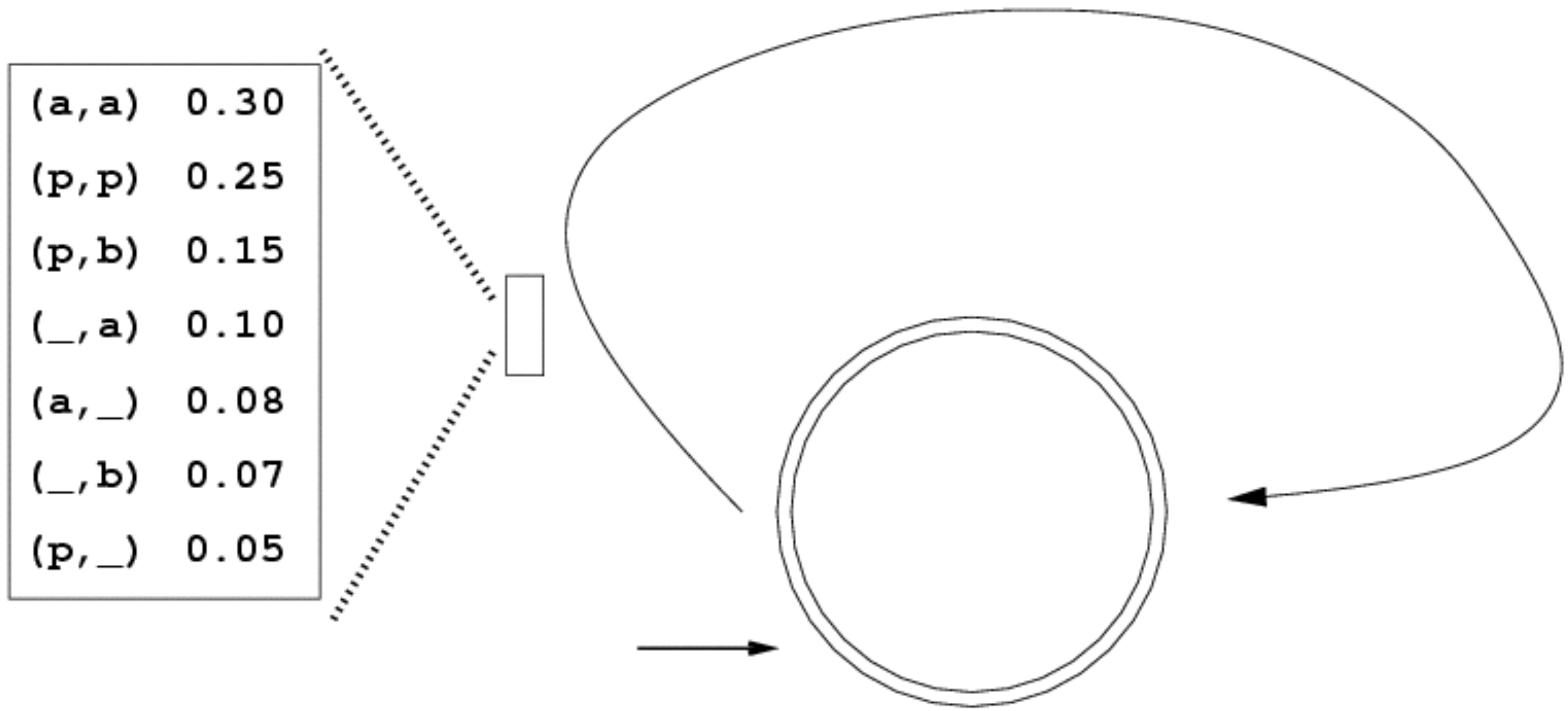
Williams: uialims uilialums uiliammas viliams

Campbell: kaampu kaampul kamvul kaamvul

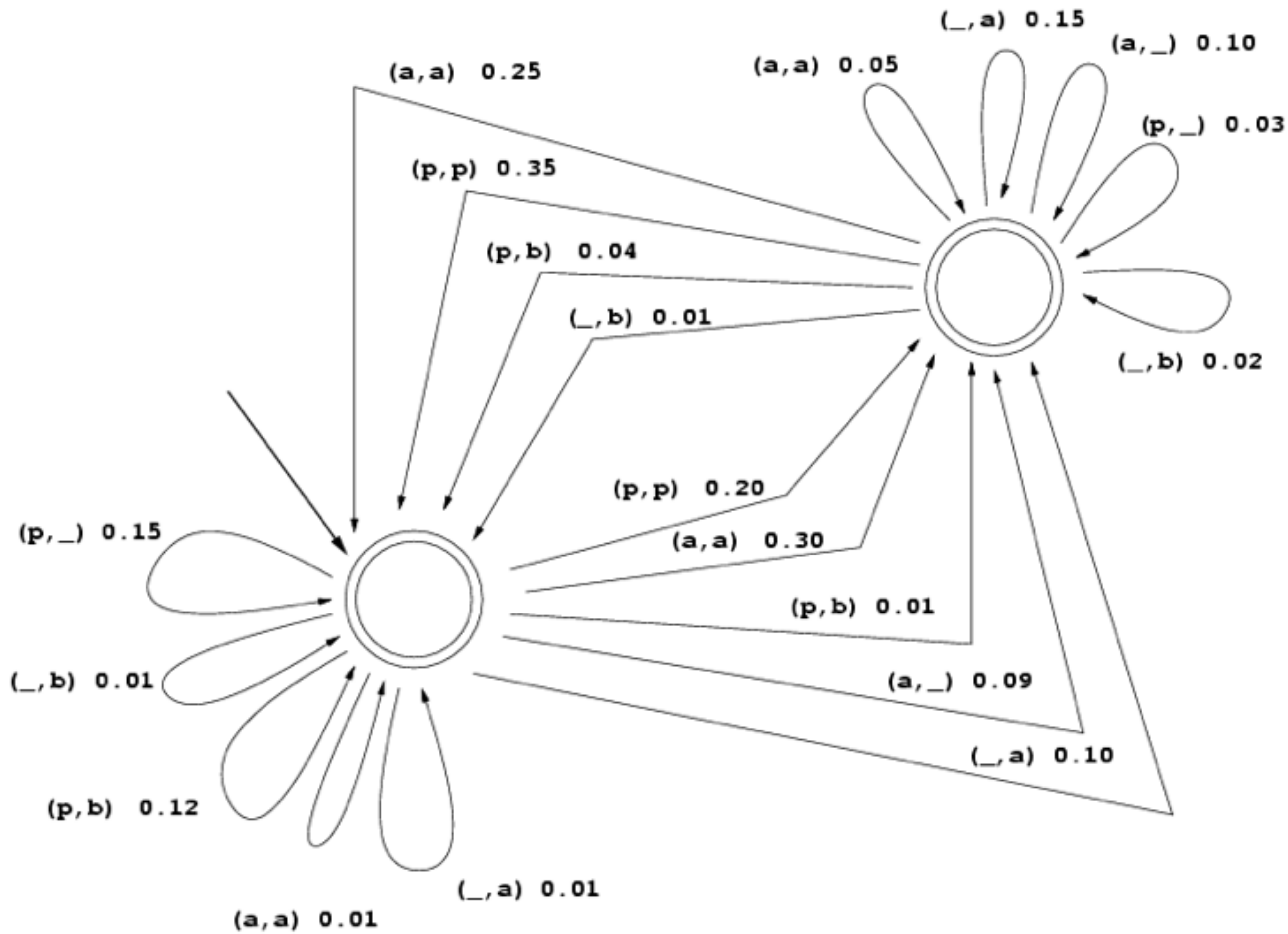
McLean: makalain maklainn makliin makkalain

Memoryless Transducer

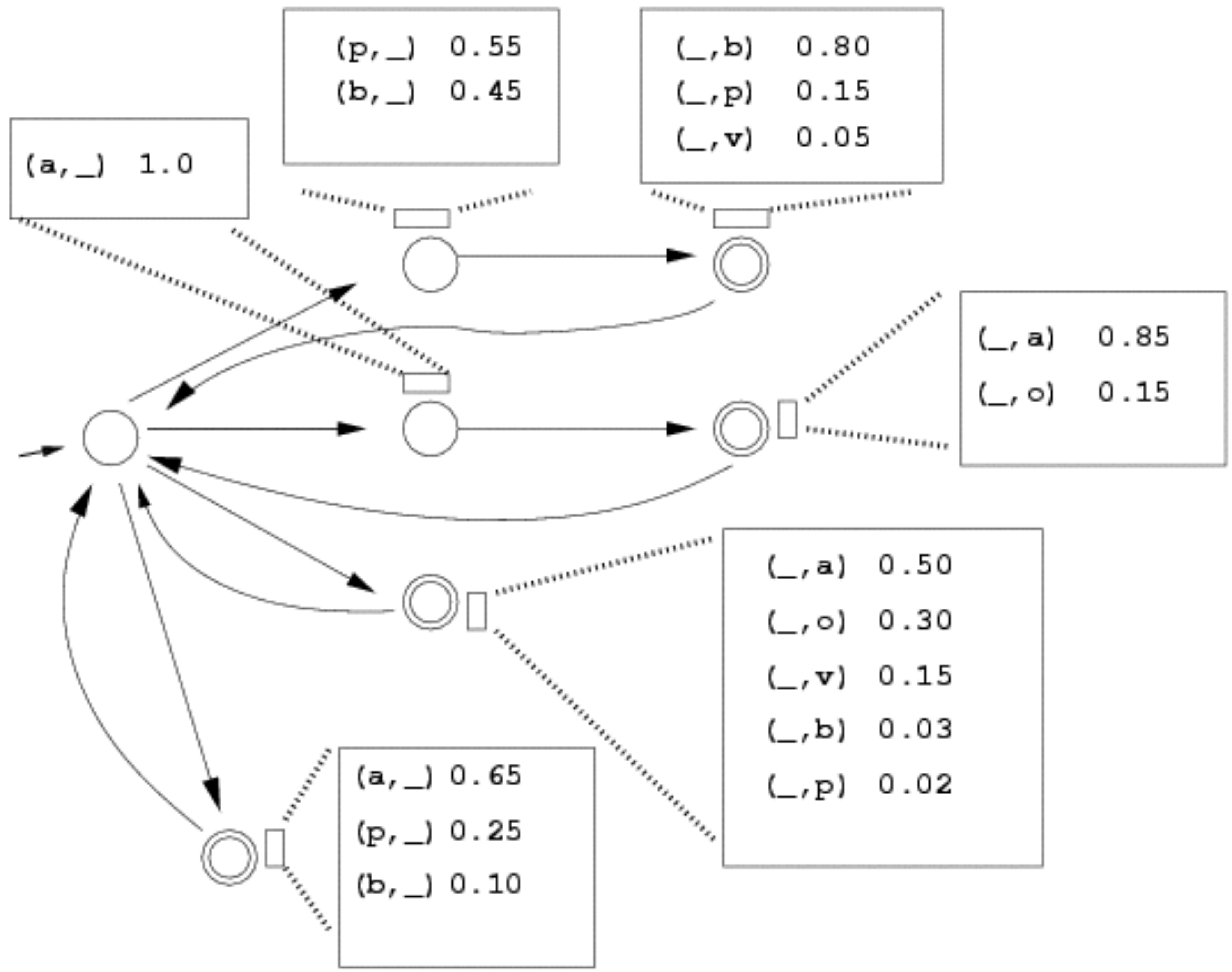
(Ristad & Yianilos 1997)



Two-State Transducer ("Weak Memory")



Unigram Interlingua Transducer



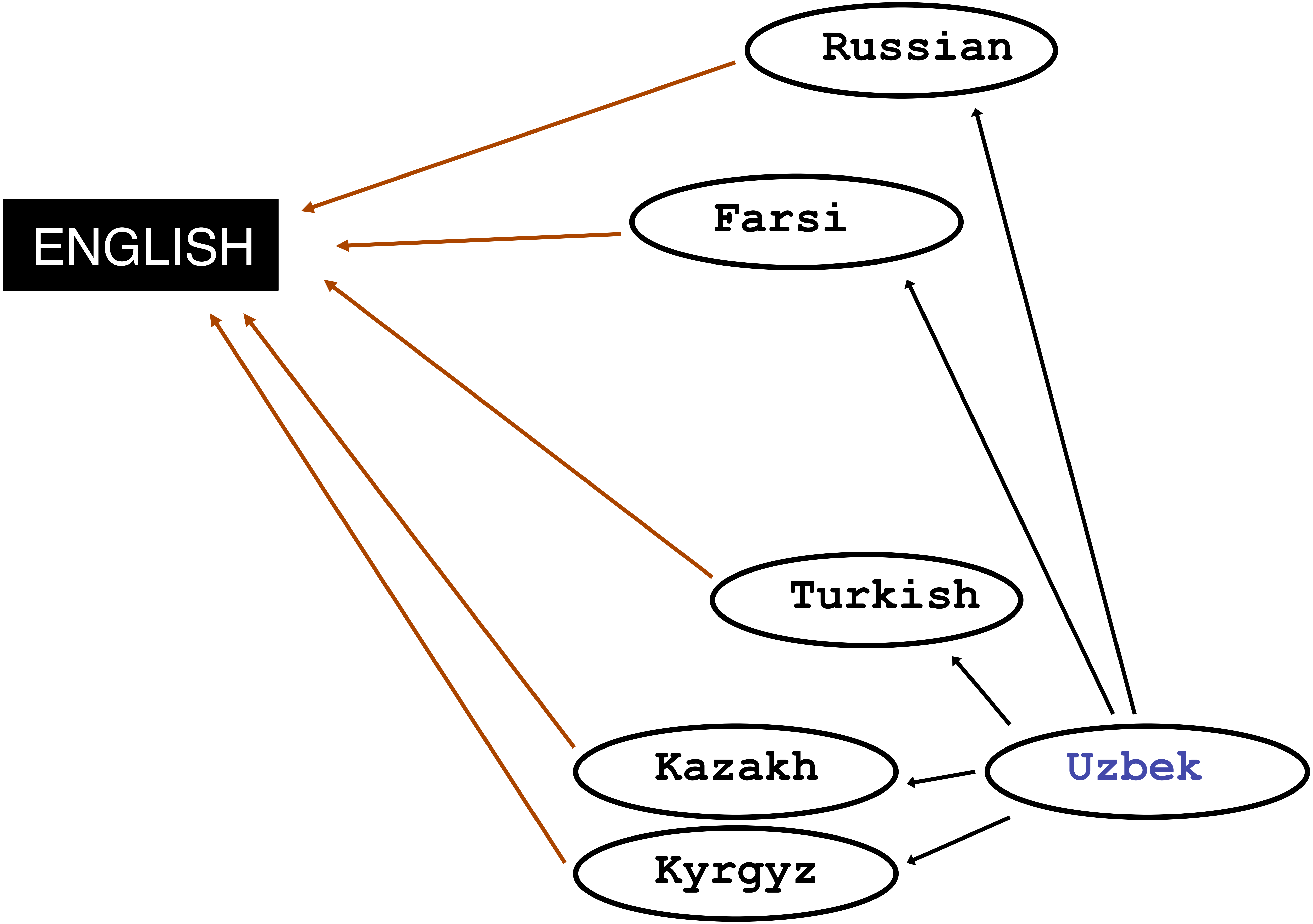
Examples: Possible Cognates Ranked by Various String Models

String Transduction Models Ranking Spanish Bridge Words for Romanian Source Word *inghiti*									
C1	**C2**	**C3**	**R&Y**	**2STEF**	**UIT**	**SN**	**AI**	**CDUI**	**JDCO**
S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato
S:ingerir	**S:ingerir**	S:engaste	S:grito	S:negrito	**S:ingerir**	S:ingente	S:negrito	S:infarto	S:engaste
S:engaste	S:engaste	**S:ingerir**	S:gaita	S:grito	S:grito	**S:ingerir**	S:negrita	S:engaste	S:anguila
S:ingreso	S:ingreso	S:inglete	S:grita	**S:ingerir**	S:grita	S:ingle	**S:ingerir**	S:ingreso	S:infarto
S:ingerido	S:ingerido	S:ingreso	S:negrito	S:negrita	S:inglete	S:angra	S:grito	S:introito	S:aguita
S:inglete	S:grito	S:ingerido	S:infarto	S:grita	S:gaita	S:ingerido	S:grita	S:negrito	S:ingreso
S:grito	S:inglete	S:infarto	S:negrita	S:gaita	S:negrito	S:ingenio	S:gaita	S:ingerido	S:intriga
S:infarto	S:infarto	S:grito	**S:ingerir**	S:ingerido	S:infarto	S:engan	S:ingenito	S:negrita	S:intuir
S:grita	S:negrito	S:introito	S:engaste	S:ingreso	S:introito	S:engatado	S:inglete	**S:ingerir**	S:indulto
S:introito	S:grita	S:engreir	S:haiti	S:haiti	S:engreir	S:invita	S:tahiti	S:inglete	S:inglete

String Transduction Models Ranking Turkish Bridge Words for Uzbek Source Word аввалги									
C1	**C2**	**C3**	**R&Y**	**2STEF**	**UIT**	**SN**	**AI**	**CDUI**	**JDCO**
T:evvelki	**T:evvelki**	**T:evvelki**	**T:evvelki**	T:vali	**T:evvelki**	T:edilgi	**T:evvelki**	**T:evvelki**	**T:evvelki**
T:evvelce	**T:evvelce**	**T:evvelce**	T:evveli	T:veli	**T:evvelce**	T:dalga	T:evveli	**T:evvelce**	**T:evvelce**
T:kalga	**T:evvelkí**	T:kalga	T:evvela	T:vals	T:edilgi	T:delgi	T:aval	T:evveli	**T:evvelkí**
T:evvelkí	T:kalga	T:salgi	**T:evvel**	T:delgi	T:algi	T:kalga	T:algi	T:evvela	T:ilkelci
T:vals	T:salgi	T:vals	T:algi	**T:evvelki**	T:salgi	T:evel	**T:evvel**	T:ilkelci	T:sivilce
T:salgi	T:vals	**T:evvelkí**	**T:evvelce**	T:kalga	T:vals	T:dalgl	T:evvela	T:eksilti	T:ilkelce
T:villa	T:villa	T:delgi	T:edilgi	T:dalga	T:delgi	**T:evvelki**	T:salgi	T:zavalli	T:akilci
T:silgi	T:silgi	T:villa	T:aval	T:villa	T:silgi	T:evlat	T:vali	**T:evvelkí**	T:eksilti
T:edilgi	T:ilkelci	T:evveli	T:evel	T:vale	T:kalga	T:dolgu	**T:evvelce**	**T:evvel**	T:asilce
T:volta	T:akilci	T:silgi	T:delgi	T:yilgi	T:dalga	T:veli	**T:evvelkí**	T:ilkelce	T:otelci

Romanian *inghiti* (ingest)
Uzbek *avvalgi* (previous/former)

* Effectiveness of cognate models



* Multi-family bridge languages

Similarity Measures

for re-ranking cognate/transliteration hypotheses

1. Probabilistic string transducers

2. Context similarity

3. Date distribution similarity

4. Similarities based on monolingual word properties

Similarity Measures

1. Probabilistic string transducers

2. Context similarity

3. Date distribution similarity

4. Similarities based on monolingual word properties

Compare Vectors

nezavisnost vector
Projection of context vector from Serbian to English term space

expression	religion	justice	majesty	sovereignty	declaration	country	ornamental
0	0	2	1.5	1.5	1.5	4	1.5

independence vector
Construction of context term vector

expression	religion	justice	majesty	sovereignty	declaration	country	ornamental
3	1	10	0	479	836	191	0

freedom vector
Construction of context term vector

expression	religion	justice	majesty	sovereignty	declaration	country	ornamental
681	184	104	0	21	4	141	0

Compute cosine similarity between nezavisnost and "independence"

… and between nezavisnost and "freedom"

Similarity Measures

1. Probabilistic string transducers

2. Context similarity

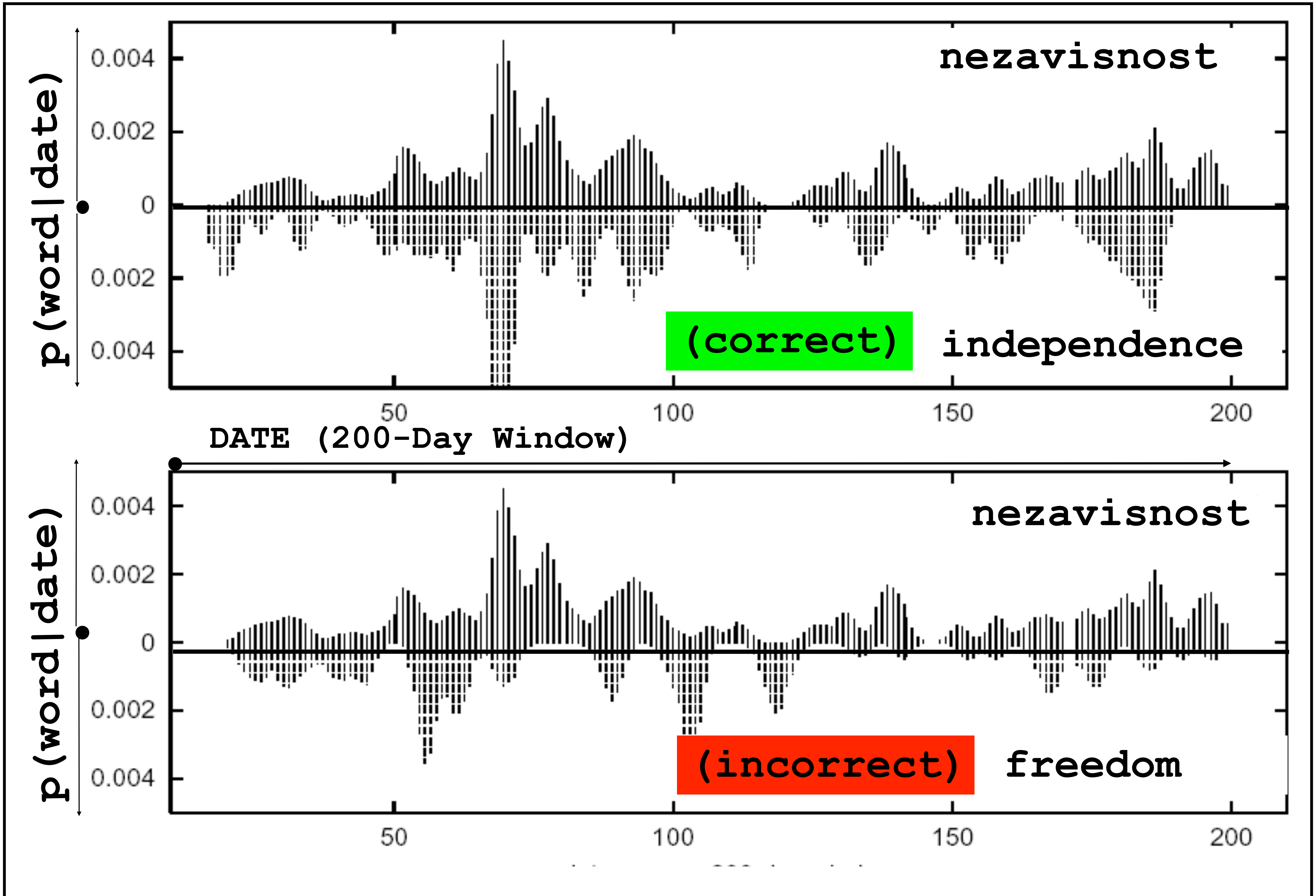
3. Date distribution similarity

4. Similarities based on monolingual word properties

Date Distribution Similarity

- Topical words associated with real-world events appear within news articles in bursts following the date of the event
- Synonymous topical words in different languages, then, display similar distributions across dates in news text: this can be measured
- We use cosine similarity on date term vectors, with term values `p(word|date)`, to quantify this notion of similarity

Date Distribution Similarity - Example



Similarity Measures

1. Probabilistic string transducers

2. Context similarity

3. Date distribution similarity

4. Similarities based on monolingual word properties

Relative Frequency

$$rf(w_F) = \frac{f_{C_F}(w_F)}{|C_F|}$$

$$rf(w_E) = \frac{f_{C_E}(w_E)}{|C_E|}$$

Cross-Language Comparison:

$$\min\left(\frac{rf(w_F)}{rf(w_E)} , \frac{rf(w_E)}{rf(w_F)} \right)$$

[min-ratio method]

Precedent in Yarowsky & Wicentowski (2000);
used relative frequency similarity for
morphological analysis

Combining Similarities: Uzbek

Individual Bridge Language Results For Uzbek Using Combined Similarity Measures				
Rank	**Turkish**	**Russian**	**Farsi**	**Kyrgyz**
1	0.04	**0.12**	0.03	0.06
5	0.10	**0.23**	0.05	0.08
10	0.13	**0.26**	0.07	0.10
20	0.16	**0.28**	0.08	0.11
50	0.21	**0.30**	0.12	0.13
100	0.24	**0.31**	0.15	0.16
200	0.26	**0.32**	0.19	0.19

Multiple Bridge Language Results For Uzbek Using Combined Similarity Measures					
Rank	**Tur+Rus**	**Tur+Rus +Farsi**	**Tur+Rus +Eng**	**Tur+Rus +Farsi +Kaz+Kyr**	**Tur+Rus +Farsi +Kaz+Kyr+Eng**
1	0.12	0.13	0.13	**0.14**	**0.14**
5	0.26	0.27	0.26	0.28	**0.29**
10	0.30	0.31	0.31	**0.34**	**0.34**
20	0.35	0.37	0.35	**0.39**	**0.39**
50	0.39	0.41	0.39	0.42	**0.43**
100	0.41	0.43	0.41	**0.46**	0.45
200	0.43	0.45	0.42	**0.48**	0.46

Combining Similarities: Romanian, Serbian, & Bengali

Multiple Bridge Language Results For Romanian Using Combined Similarity Measures				
Rank	Spanish	Spanish +Russian	Spanish +English	Spanish +Russian +English
1	0.17	0.18	**0.19**	**0.19**
5	0.31	0.35	0.34	**0.37**
10	0.37	0.41	0.41	**0.43**
20	0.43	0.46	0.46	**0.48**
50	0.51	0.53	0.53	**0.55**
100	0.57	0.60	0.58	**0.61**
200	0.60	**0.62**	0.59	**0.62**

Multiple Bridge Language Results For Serbian Using Combined Similarity Measures						
Rank	Cz	Rus	Bulg	Cz +English	Cz+Slovak +Rus+Bulg	Cz+Slovak +Rus+Bulg +English
1	0.13	0.15	**0.19**	0.13	**0.19**	**0.19**
5	0.24	0.24	0.31	0.25	**0.38**	**0.38**
10	0.29	0.28	0.35	0.30	0.42	**0.43**
20	0.32	0.31	0.40	0.34	**0.48**	**0.48**
50	0.38	0.36	0.44	0.39	0.54	**0.55**
100	0.40	0.40	0.48	0.42	**0.59**	**0.59**
200	0.41	0.42	0.50	0.43	**0.60**	**0.60**

Bridge Language Results for Bengali Using Combined Similarity Measures		
Rank	Hindi	Hindi +English
1	0.03	**0.05**
5	0.11	**0.14**
10	0.13	**0.17**
20	0.16	**0.21**
50	0.19	**0.25**
100	0.22	**0.28**
200	0.23	**0.29**

Observations

* With no Uzbek-specific supervision,
we can produce an Uzbek-English
dictionary which is 14% exact-match correct

* Or, we can put a correct translation
in the top-10 list 34% of the time
(useful for end-to-end machine translation
or cross-language information retrieval)

* Adding more
bridge languages
helps

Multiple Bridge Language Results For Uzbek
Using Combined Similarity Measures

Rank	Tur+Rus	Tur+Rus +Farsi	Tur+Rus +Eng	Tur+Rus +Farsi +Kaz+Kyr	Tur+Rus +Farsi +Kaz+Kyr+Eng
1	0.12	0.13	0.13	**0.14**	**0.14**
5	0.26	0.27	0.26	0.28	**0.29**
10	0.30	0.31	0.31	**0.34**	**0.34**
20	0.35	0.37	0.35	**0.39**	**0.39**
50	0.39	0.41	0.39	0.42	**0.43**
100	0.41	0.43	0.41	**0.46**	0.45
200	0.43	0.45	0.42	**0.48**	0.46

Topic Models

Text Reuse

Jobless rate at 3-year low as payrolls surge

Recommend 1,328 people recommend this.

Tweet

Share

Share t

+1 49

Email

Print

Related Ne

Instant vi
nonfarm p
rose by 2
Fri, Feb 3 20

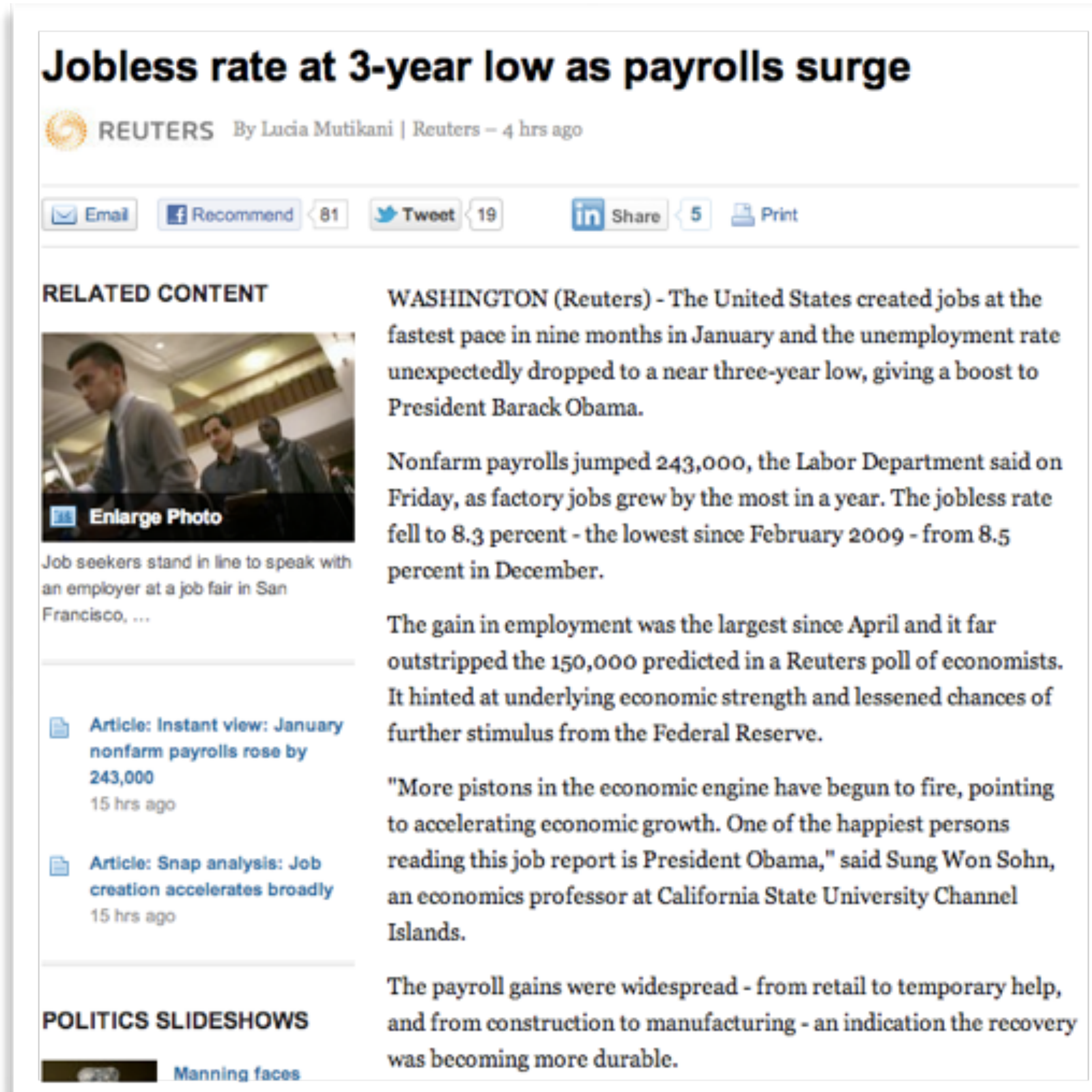
Snap ana
creation a
broadly
Fri, Feb 3 20

Analysis &

Only 45,000
created in J
TrimTabs

By Lucia Mutikani
WASHINGTON | Fri Feb 3, 2012 5:35pm EST

(Reuters) - The United States created jobs at the fastest pace in nine months in January and the unemployment rate unexpectedly dropped to a near three-year low, giving a boost to President Barack Obama.



Jobless rate at 3-year low as payrolls surge

REUTERS By Lucia Mutikani | Reuters – 4 hrs ago

Email | Recommend 81 | Tweet 19 | Share 5 | Print

RELATED CONTENT

Enlarge Photo

Job seekers stand in line to speak with an employer at a job fair in San Francisco, ...

Article: Instant view: January nonfarm payrolls rose by 243,000
15 hrs ago

Article: Snap analysis: Job creation accelerates broadly
15 hrs ago

POLITICS SLIDESHOWS

Manning faces

WASHINGTON (Reuters) - The United States created jobs at the fastest pace in nine months in January and the unemployment rate unexpectedly dropped to a near three-year low, giving a boost to President Barack Obama.

Nonfarm payrolls jumped 243,000, the Labor Department said on Friday, as factory jobs grew by the most in a year. The jobless rate fell to 8.3 percent - the lowest since February 2009 - from 8.5 percent in December.

The gain in employment was the largest since April and it far outstripped the 150,000 predicted in a Reuters poll of economists. It hinted at underlying economic strength and lessened chances of further stimulus from the Federal Reserve.

"More pistons in the economic engine have begun to fire, pointing to accelerating economic growth. One of the happiest persons reading this job report is President Obama," said Sung Won Sohn, an economics professor at California State University Channel Islands.

The payroll gains were widespread - from retail to temporary help, and from construction to manufacturing - an indication the recovery was becoming more durable.

Topical Similarity

Job Gains Reflect Hope a Recovery Is Blooming

Joan Barnett Lee/The Modesto Bee, via Associated Press

A job applicant received assistance at an employment fair in Modesto, Calif., this week.

By MOTOKO RICH
Published: February 3, 2012

The front wheels have lifted off the runway. Now, Americans are waiting to see if the economy can truly get aloft.

RECOMMEND
TWITTER
LINKEDIN
COMMENTS (576)
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

Multimedia

Jobs | Private | Rate

Change in jobs, in thousands

With the government reporting that the unemployment rate and the number of jobless fell in January to the lowest levels since early 2009, the recovery seems finally to be reaching American workers.

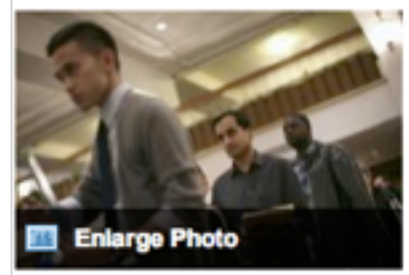
The Labor Department's latest

Jobless rate at 3-year low as payrolls surge

REUTERS By Lucia Mutikani | Reuters – 4 hrs ago

Email | Recommend 81 | Tweet 19 | Share 5 | Print

RELATED CONTENT



Job seekers stand in line to speak with an employer at a job fair in San Francisco, ...

Article: Instant view: January nonfarm payrolls rose by 243,000
15 hrs ago

Article: Snap analysis: Job creation accelerates broadly
15 hrs ago

POLITICS SLIDESHOWS

Manning faces

WASHINGTON (Reuters) - The United States created jobs at the fastest pace in nine months in January and the unemployment rate unexpectedly dropped to a near three-year low, giving a boost to President Barack Obama.

Nonfarm payrolls jumped 243,000, the Labor Department said on Friday, as factory jobs grew by the most in a year. The jobless rate fell to 8.3 percent - the lowest since February 2009 - from 8.5 percent in December.

The gain in employment was the largest since April and it far outstripped the 150,000 predicted in a Reuters poll of economists. It hinted at underlying economic strength and lessened chances of further stimulus from the Federal Reserve.

"More pistons in the economic engine have begun to fire, pointing to accelerating economic growth. One of the happiest persons reading this job report is President Obama," said Sung Won Sohn, an economics professor at California State University Channel Islands.

The payroll gains were widespread - from retail to temporary help, and from construction to manufacturing - an indication the recovery was becoming more durable.

Parallel Bitext

Genehmigung des Protokolls	Approval of the minutes
Das Protokoll der Sitzung vom Donnerstag, den 28. März 1996 wurde verteilt.	The minutes of the sitting of Thursday, 28 March 1996 have been distributed.
Gibt es Einwände?	Are there any comments?
Die Punkte 3 und 4 widersprechen sich jetzt, obwohl es bei der Abstimmung anders aussah.	Points 3 and 4 now contradict one another whereas the voting showed otherwise.
Das muß ich erst einmal klären, Frau Oomen-Ruijten.	I will have to look into that, Mrs Oomen-Ruijten.

Koehn (2005): European Parliament corpus

Multilingual Topical Similarity

Abraham Lincoln

From Wikipedia, the free encyclopedia

This article is about the American president. For other uses, see Abraham Lincoln (disambiguation).

Abraham Lincoln /ˈeɪbrəhæm ˈlɪŋkən/ (February 12, 1809 – April 15, 1865) was the 16th President of the United States, serving from March 1861 until his assassination in April 1865. He successfully led his country through a great constitutional, military and moral crisis – the American Civil War – preserving the Union, while ending slavery, and promoting economic and financial modernization. Reared in a poor family on the western frontier, Lincoln was mostly self-educated. He became a country lawyer, an Illinois state legislator, and a one-term member of the United States House of Representatives, but failed in two attempts to be elected to the United States Senate.

Abraham Lincoln

Abraham Lincoln [ˈeɪbrəhæm ˈliŋkən] (* 12. Februar 1809 bei Hodgenville, Hardin County, heute: LaRue County, Kentucky; † 15. April 1865 in Washington, D.C.) amtierte von 1861 bis 1865 als 16. Präsident der Vereinigten Staaten von Amerika. Er war der erste aus den Reihen der Republikanischen Partei und der erste, der einem Attentat zum Opfer fiel. 1860 gewählt, gelang ihm 1864 die Wiederwahl.

Seine Präsidentschaft gilt als eine der bedeutendsten in der Geschichte der Vereinigten Staaten: Die Wahl des Sklavereigegners veranlasste zunächst sieben, später weitere vier der sklavenhaltenden Südstaaten zur Sezession. Lincoln führte die verbliebenen Nordstaaten durch den daraus entstandenen Bürgerkrieg, setzte die Wiederherstellung der Union durch und betrieb erfolgreich die Abschaffung der Sklaverei in den Vereinigten Staaten. Unter seiner Regierung schlugen die USA den Weg zum zentral regierten, modernen Industriestaat ein und schufen so die Basis für ihren Aufstieg zur Weltmacht im 20. Jahrhundert.

What Representation?

What Representation?

- Bag of words, n-grams, etc.?

What Representation?

- Bag of words, n-grams, etc.?
  - Vocabulary mismatch within language:

What Representation?

- Bag of words, n-grams, etc.?
  - Vocabulary mismatch within language:
    - *Jobless* vs. *unemployed*

What Representation?

- Bag of words, n-grams, etc.?
  - Vocabulary mismatch within language:
    - *Jobless* vs. *unemployed*
  - What about between languages?

What Representation?

- Bag of words, n-grams, etc.?
  - Vocabulary mismatch within language:
    - *Jobless* vs. *unemployed*
  - What about between languages?
    - Translate everything into English?

What Representation?

- Bag of words, n-grams, etc.?
  - Vocabulary mismatch within language:
    - *Jobless* vs. *unemployed*
  - What about between languages?
    - Translate everything into English?
- Represent documents/passages as probability distributions over hidden “topics”

Plate Notation

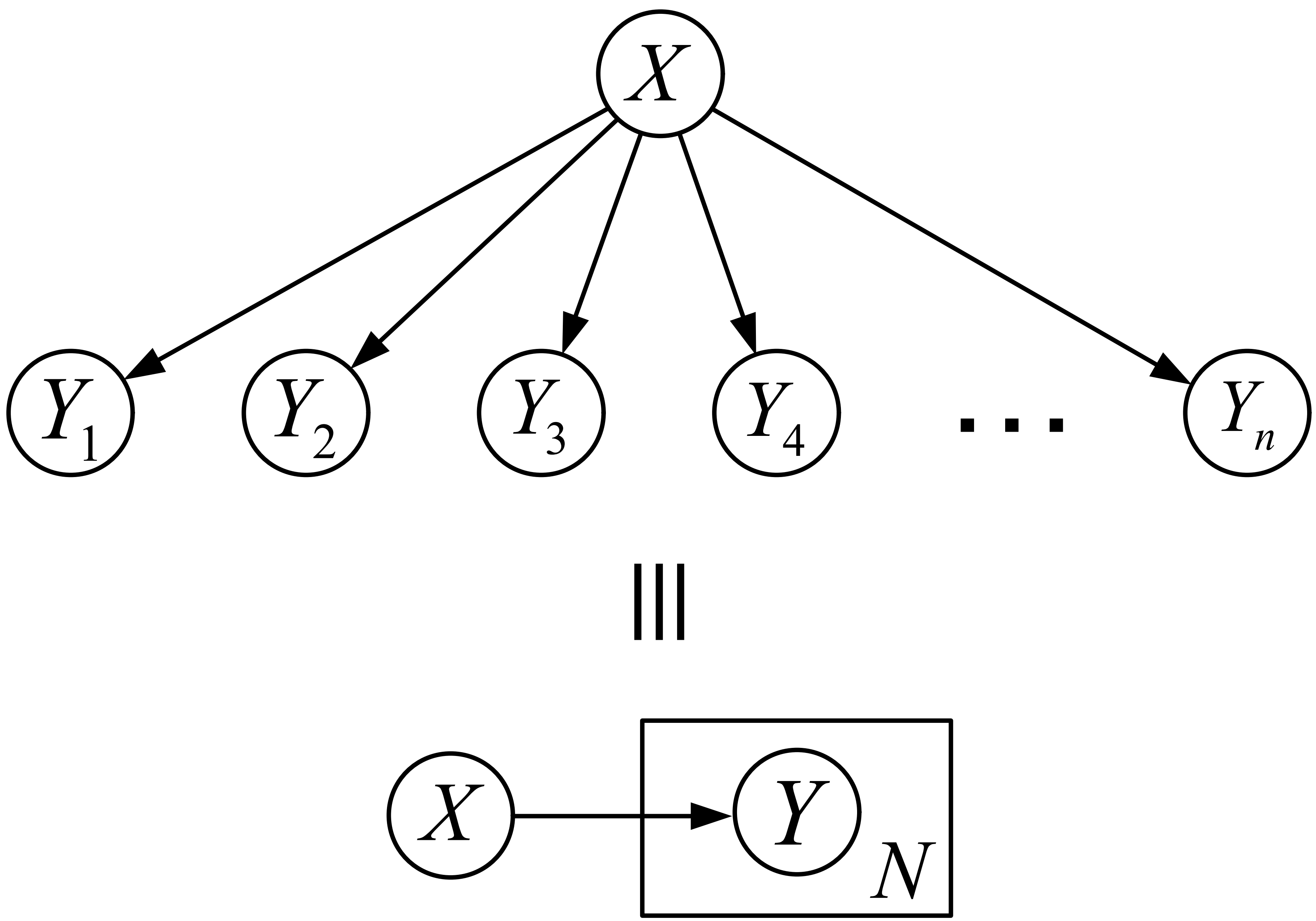
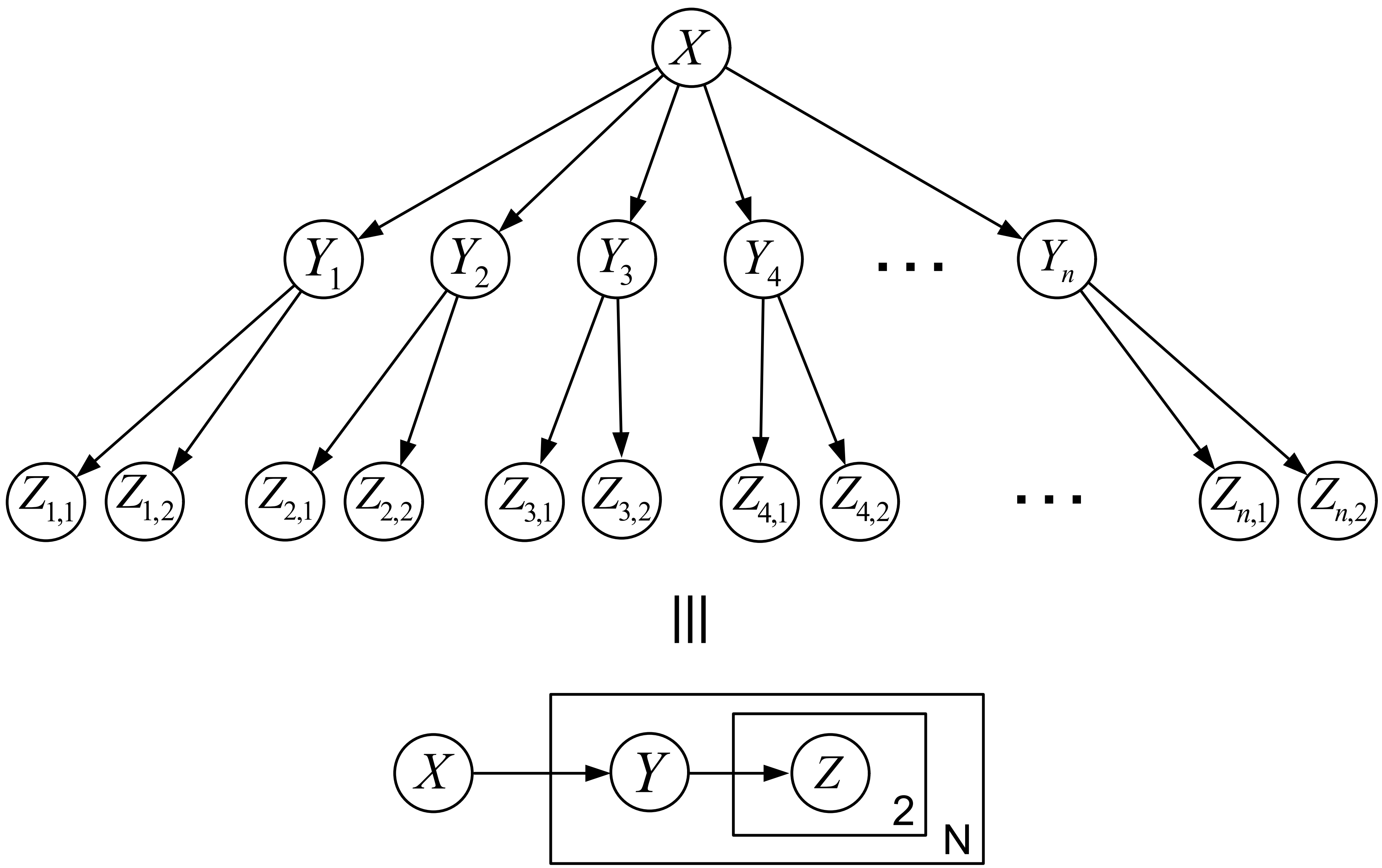
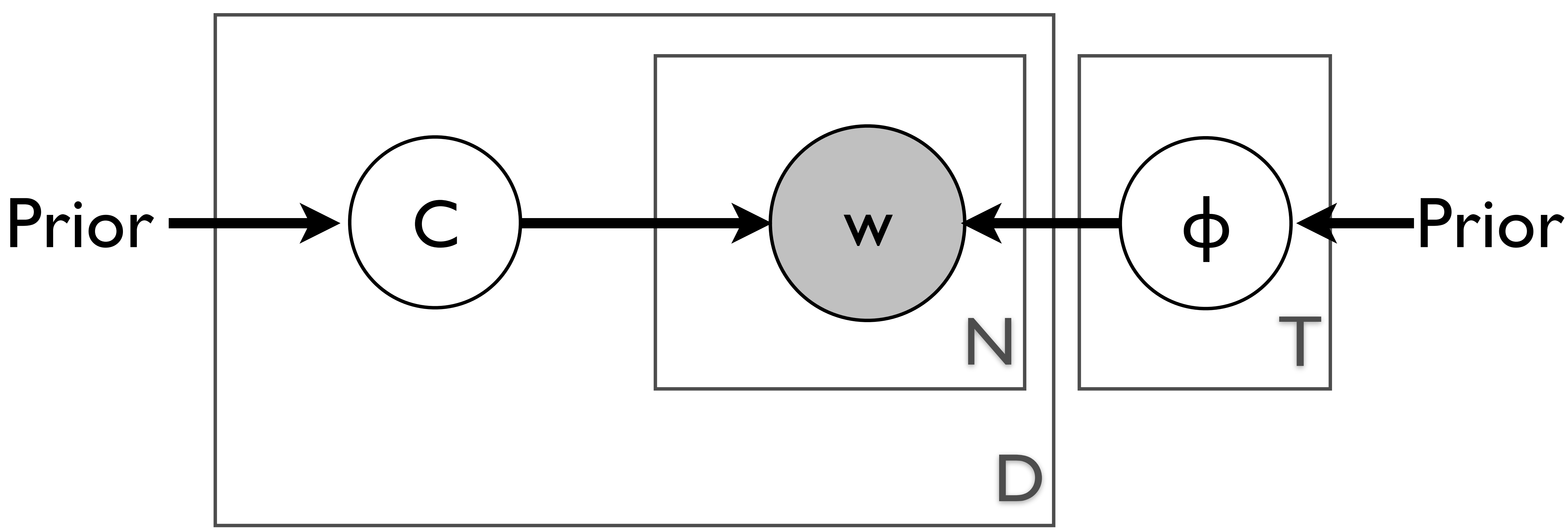


Plate Notation



Modeling Text with Naive Bayes

- Let the text talk about *T* topics
- Each topic is a probability dist'n over all words
- For *D* documents each with N_D words:



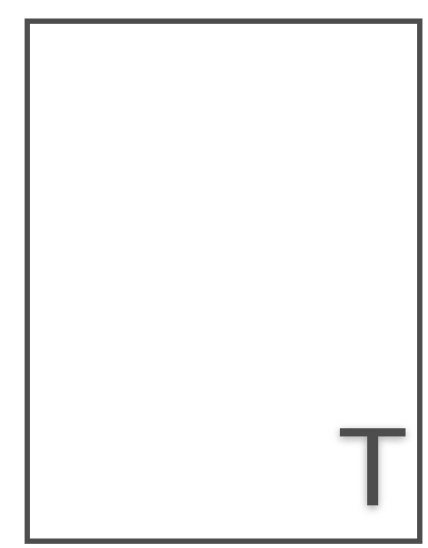
Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

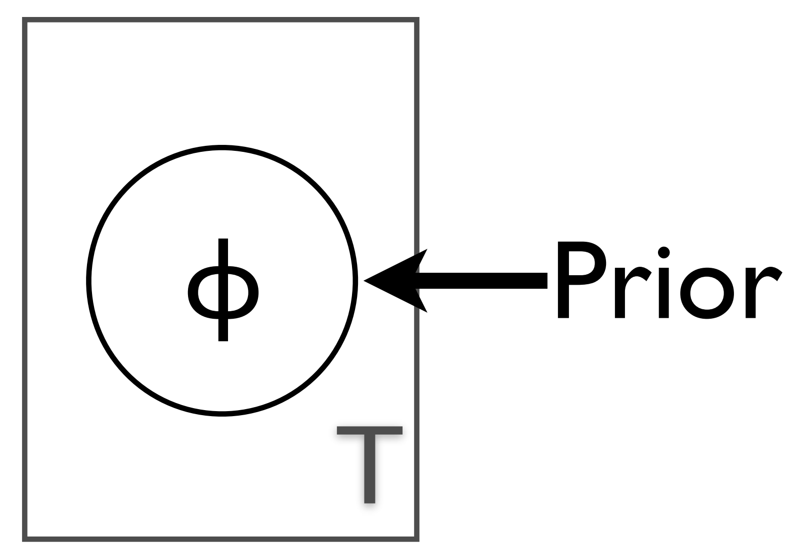
- Let the text talk about *T* topics



Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

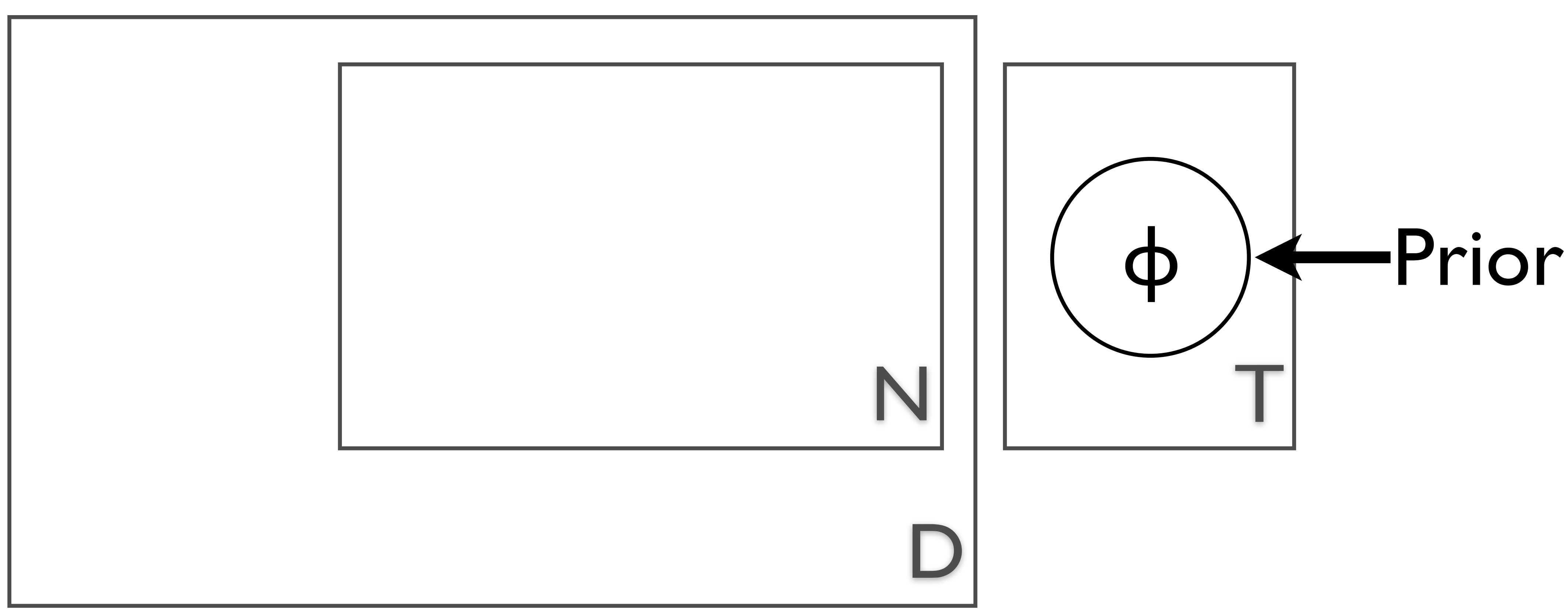
- Let the text talk about *T* topics
- Each topic is a probability dist'n over all words



Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

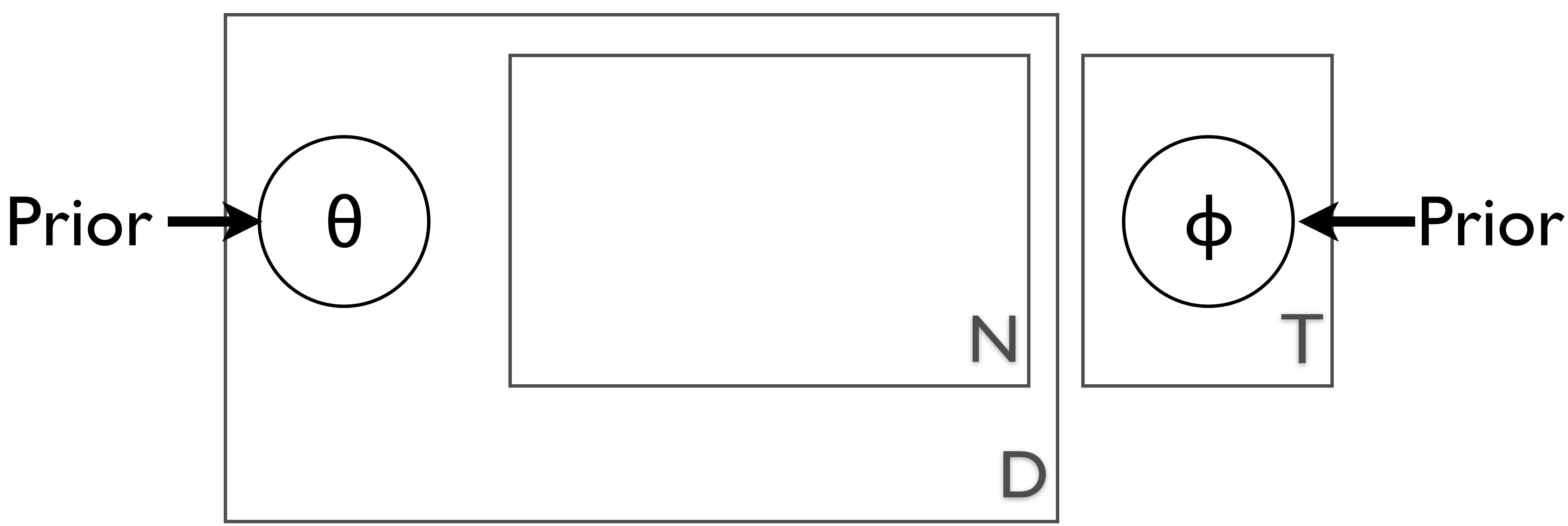
- Let the text talk about *T* topics
- Each topic is a probability dist'n over all words
- For *D* documents each with N_D words:



Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

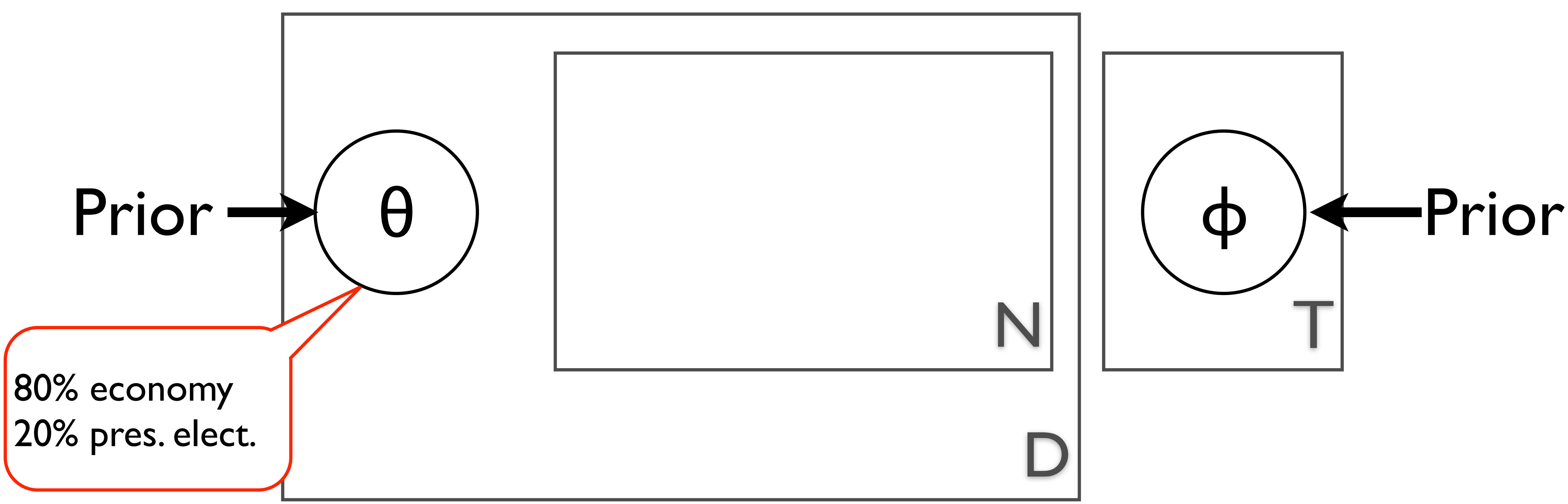
- Let the text talk about *T* topics
- Each topic is a probability dist’n over all words
- For *D* documents each with N_D words:



Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

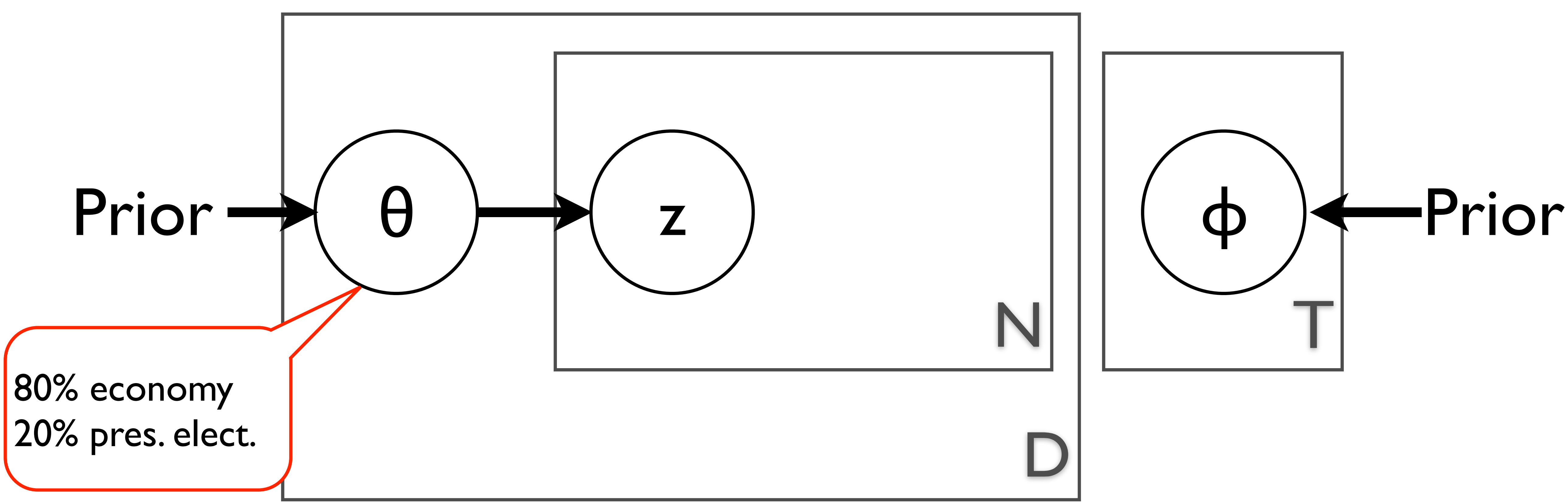
- Let the text talk about *T* topics
- Each topic is a probability dist'n over all words
- For *D* documents each with N_D words:



Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

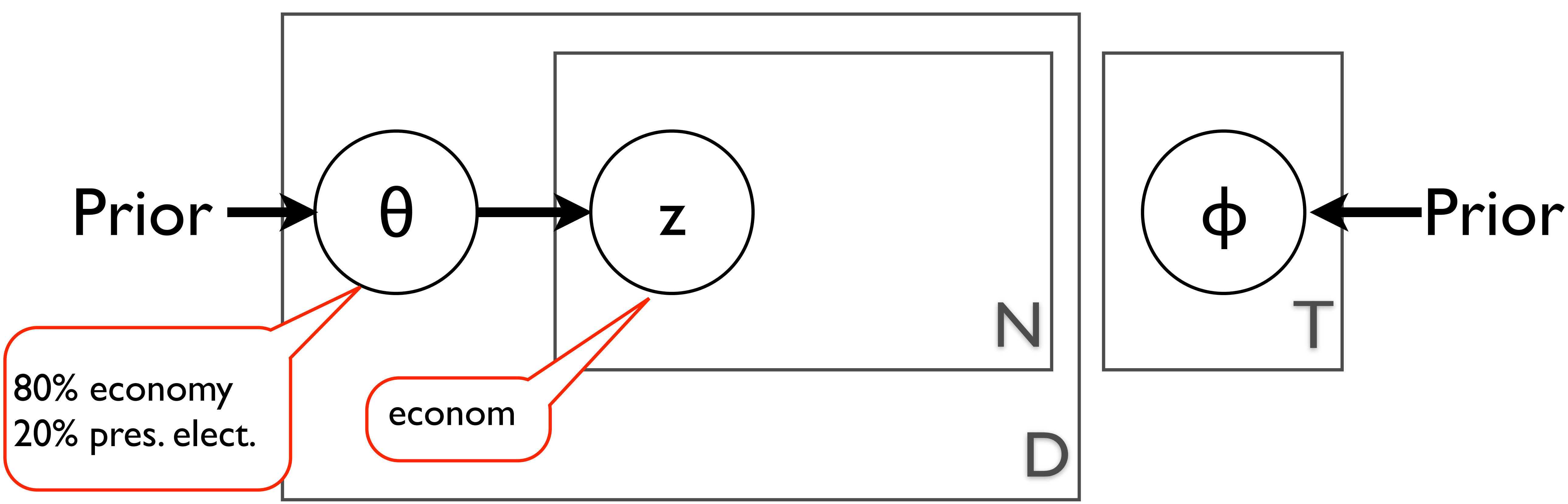
- Let the text talk about *T* topics
- Each topic is a probability dist'n over all words
- For *D* documents each with N_D words:



Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

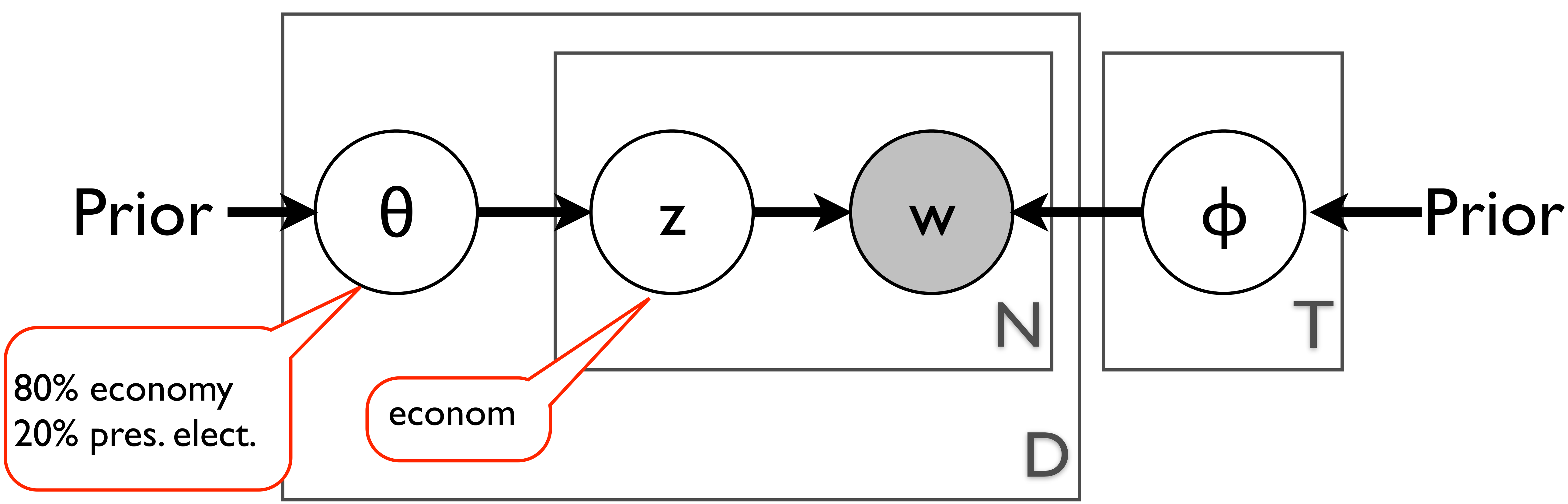
- Let the text talk about T topics
- Each topic is a probability dist'n over all words
- For D documents each with N_D words:



Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

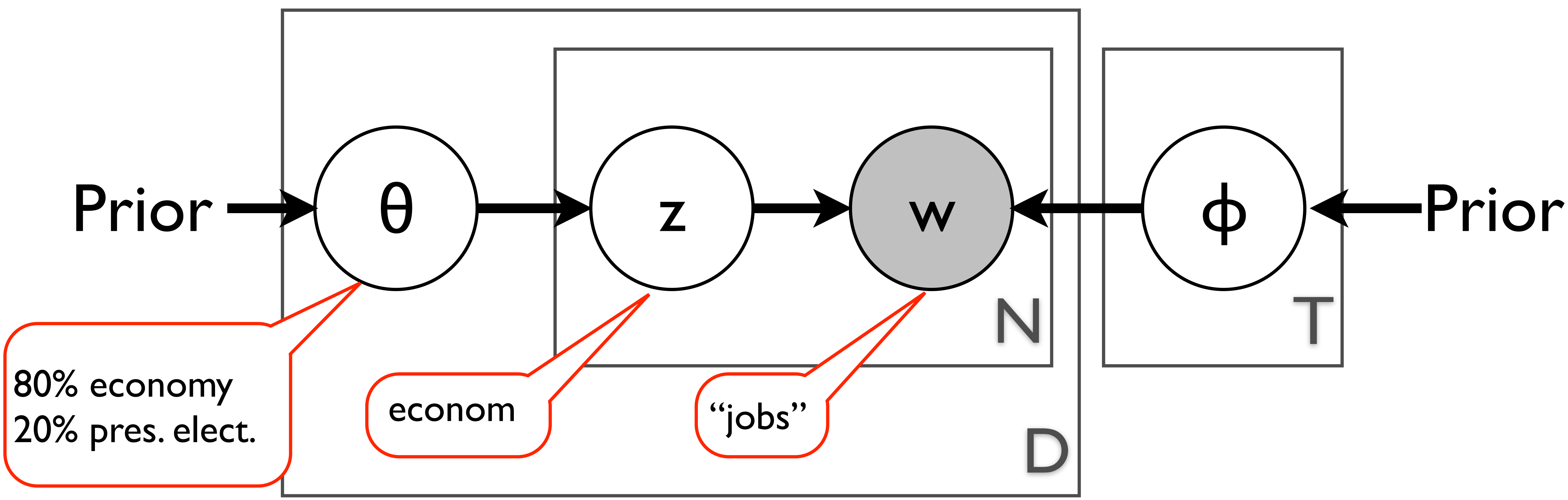
- Let the text talk about *T* topics
- Each topic is a probability dist'n over all words
- For *D* documents each with N_D words:

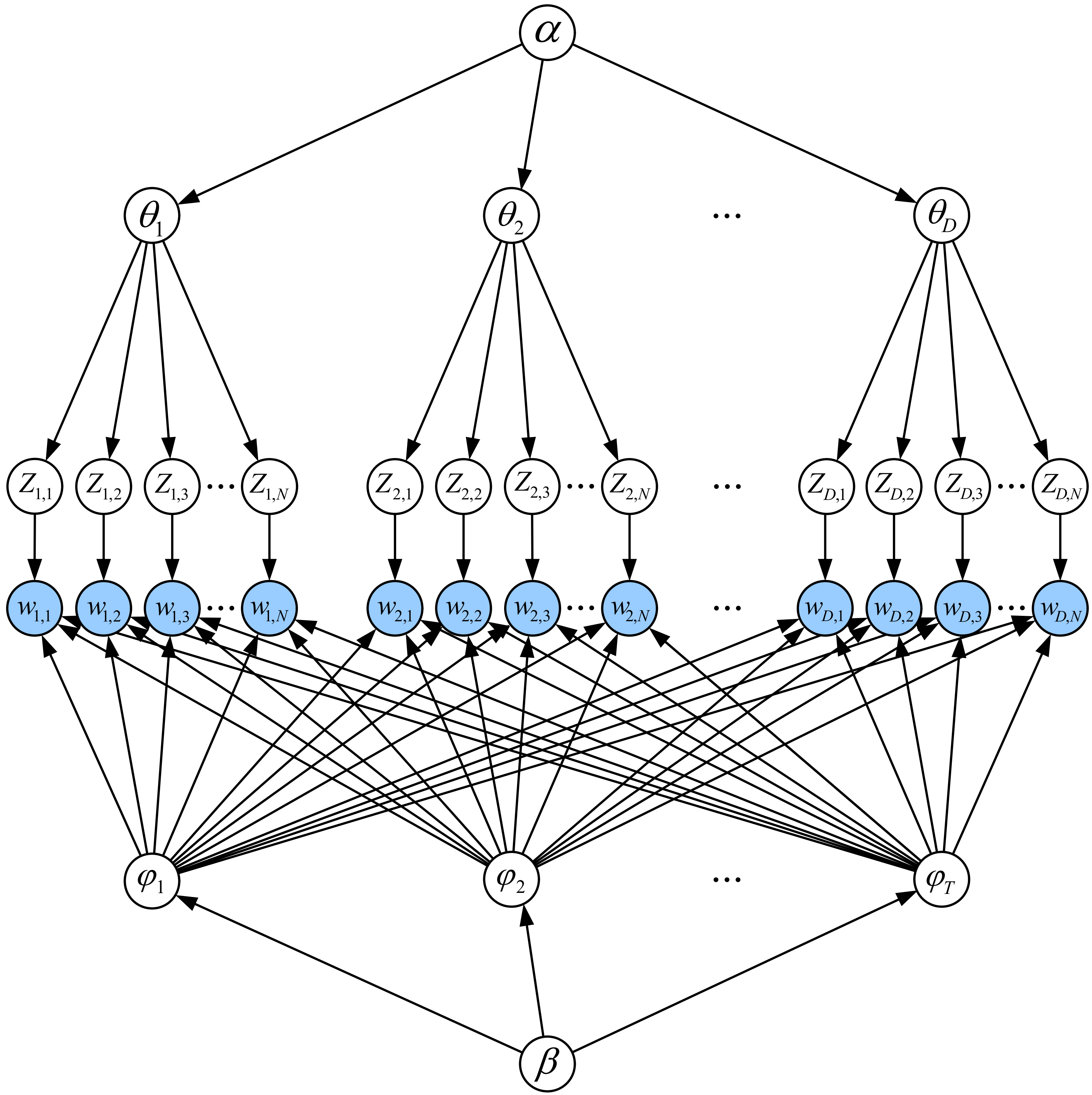


Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

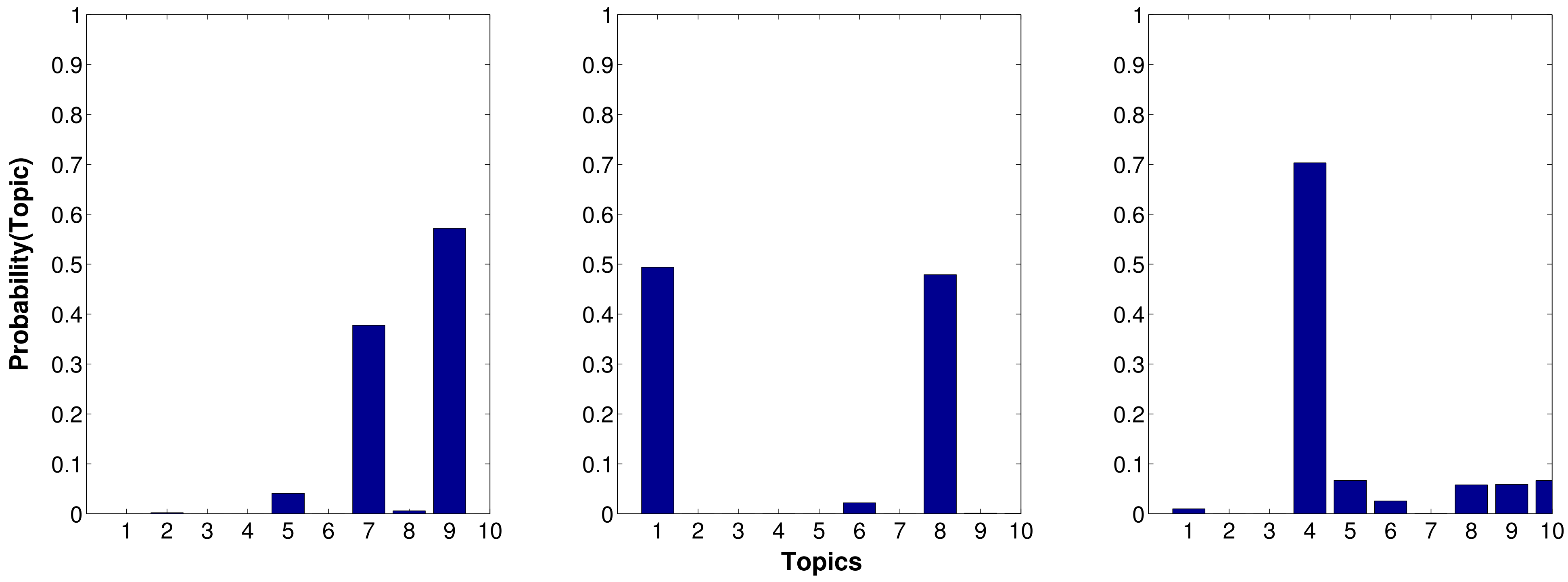
- Let the text talk about *T* topics
- Each topic is a probability dist'n over all words
- For *D* documents each with N_D words:



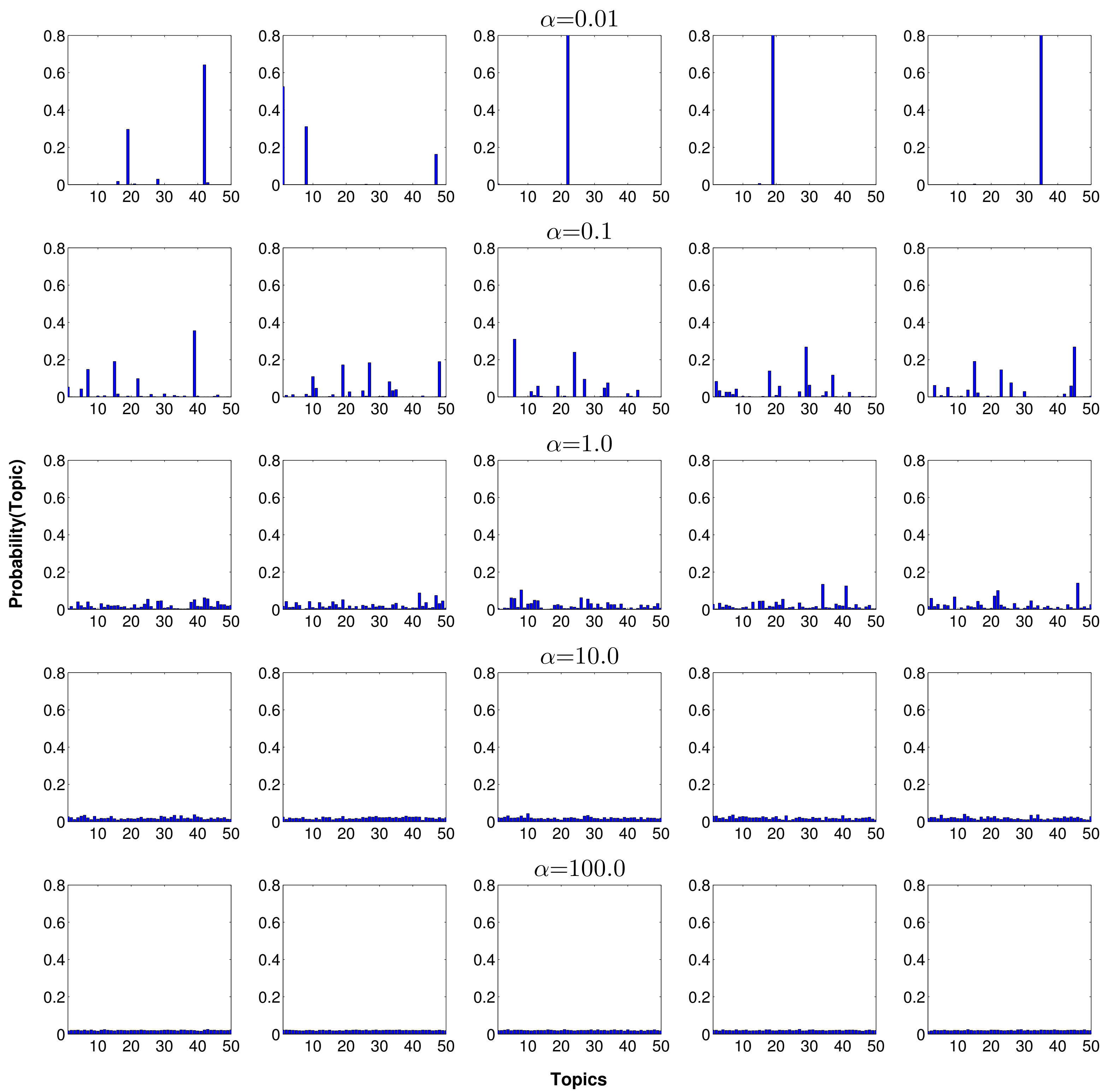


α
θ1
θ2
θD
Z1,1
Z1,2
Z1,3
Z1,N
Z2,1
Z2,2
Z2,3
Z2,N
ZD,1
ZD,2
ZD,3
ZD,N
w1,1
w1,2
w1,3
w1,N
w2,1
w2,2
w2,3
w2,N
wD,1
wD,2
wD,3
wD,N
φ1
φ2
φT
β

Multinomials as Histograms



Dirichlet Priors on Histograms



Top Words by Topic

Topics →

1	2	3	4	5	6	7	8
DISEASE	WATER	MIND	STORY	FIELD	SCIENCE	BALL	JOB
BACTERIA	FISH	WORLD	STORIES	MAGNETIC	STUDY	GAME	WORK
DISEASES	SEA	DREAM	TELL	MAGNET	SCIENTISTS	TEAM	JOBS
GERMS	SWIM	DREAMS	CHARACTER	WIRE	SCIENTIFIC	FOOTBALL	CAREER
FEVER	SWIMMING	THOUGHT	CHARACTERS	NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CAUSE	POOL	IMAGINATION	AUTHOR	CURRENT	WORK	PLAYERS	EMPLOYMENT
CAUSED	LIKE	MOMENT	READ	COIL	RESEARCH	PLAY	OPPORTUNITIES
SPREAD	SHELL	THOUGHTS	TOLD	POLES	CHEMISTRY	FIELD	WORKING
VIRUSES	SHARK	OWN	SETTING	IRON	TECHNOLOGY	PLAYER	TRAINING
INFECTION	TANK	REAL	TALES	COMPASS	MANY	BASKETBALL	SKILLS
VIRUS	SHELLS	LIFE	PLOT	LINES	MATHEMATICS	COACH	CAREERS
MICROORGANISMS	SHARKS	IMAGINE	TELLING	CORE	BIOLOGY	PLAYED	POSITIONS
PERSON	DIVING	SENSE	SHORT	ELECTRIC	FIELD	PLAYING	FIND
INFECTIOUS	DOLPHINS	CONSCIOUSNESS	FICTION	DIRECTION	PHYSICS	HIT	POSITION
COMMON	SWAM	STRANGE	ACTION	FORCE	LABORATORY	TENNIS	FIELD
CAUSING	LONG	FEELING	TRUE	MAGNETS	STUDIES	TEAMS	OCCUPATIONS
SMALLPOX	SEAL	WHOLE	EVENTS	BE	WORLD	GAMES	REQUIRE
BODY	DIVE	BEING	TELLS	MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
INFECTIONS	DOLPHIN	MIGHT	TALE	POLE	STUDYING	BAT	EARN
CERTAIN	UNDERWATER	HOPE	NOVEL	INDUCED	SCIENCES	TERRY	ABLE

Griffiths et al.

Top Words by Topic

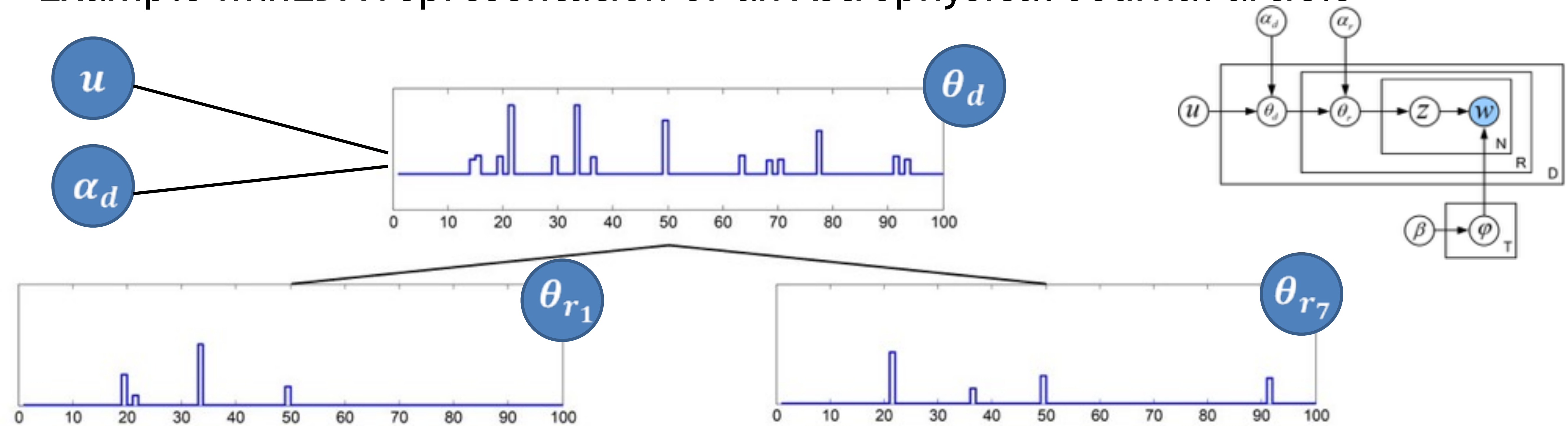
Topics →

1	2	3	4	5	6	7	8
DISEASE	WATER	MIND	STORY	**FIELD**	SCIENCE	BALL	JOB
BACTERIA	FISH	WORLD	STORIES	MAGNETIC	STUDY	GAME	WORK
DISEASES	SEA	DREAM	TELL	MAGNET	SCIENTISTS	TEAM	JOBS
GERMS	SWIM	DREAMS	CHARACTER	WIRE	SCIENTIFIC	FOOTBALL	CAREER
FEVER	SWIMMING	THOUGHT	CHARACTERS	NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CAUSE	POOL	IMAGINATION	AUTHOR	CURRENT	WORK	PLAYERS	EMPLOYMENT
CAUSED	LIKE	MOMENT	READ	COIL	RESEARCH	PLAY	OPPORTUNITIES
SPREAD	SHELL	THOUGHTS	TOLD	POLES	CHEMISTRY	**FIELD**	WORKING
VIRUSES	SHARK	OWN	SETTING	IRON	TECHNOLOGY	PLAYER	TRAINING
INFECTION	TANK	REAL	TALES	COMPASS	MANY	BASKETBALL	SKILLS
VIRUS	SHELLS	LIFE	PLOT	LINES	MATHEMATICS	COACH	CAREERS
MICROORGANISMS	SHARKS	IMAGINE	TELLING	CORE	BIOLOGY	PLAYED	POSITIONS
PERSON	DIVING	SENSE	SHORT	ELECTRIC	**FIELD**	PLAYING	FIND
INFECTIOUS	DOLPHINS	CONSCIOUSNESS	FICTION	DIRECTION	PHYSICS	HIT	POSITION
COMMON	SWAM	STRANGE	ACTION	FORCE	LABORATORY	TENNIS	**FIELD**
CAUSING	LONG	FEELING	TRUE	MAGNETS	STUDIES	TEAMS	OCCUPATIONS
SMALLPOX	SEAL	WHOLE	EVENTS	BE	WORLD	GAMES	REQUIRE
BODY	DIVE	BEING	TELLS	MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
INFECTIONS	DOLPHIN	MIGHT	TALE	POLE	STUDYING	BAT	EARN
CERTAIN	UNDERWATER	HOPE	NOVEL	INDUCED	SCIENCES	TERRY	ABLE

Griffiths et al.

Hierarchical Document Models

- Example mlhLDA representation of an Astrophysical Journal article



1. INTRODUCTION
Blazars are an intriguing class of active galactic nuclei (AGNs), dominated by non-thermal radiation over the entire electromagnetic spectrum. Their emission extends from radio to TeV energies with a broadband spectral energy distribution (SED) typically described by two main components, the first peaking from IR to X=ray energy range in which blazars are the most commonly detected extragalactic sources...

Rank	Topic = 32	Topic = 48	Topic = 18
1	spectral	measured	aperture
2	amplification	uncertainties	measured
3	isotropic	catalog	total
4	dropout	matching	exposure
5	competition	estimated	position
6	caustic	respectively	ratio
7	detected	final	selected
8	antenna	cathode	color
9	function	total	spread
10	color	limit	objects

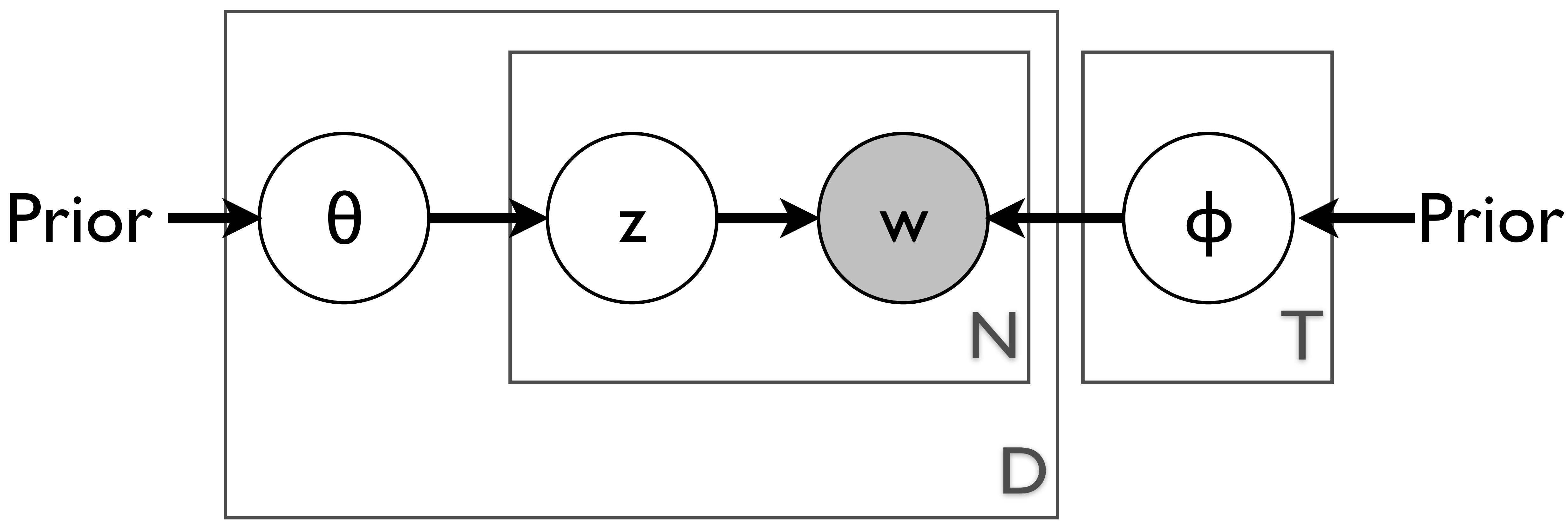
...

7. SUMMARY AND DISCUSSION
We have presented the infrared characterization of a sample of blazars detected in the γ-ray. In order to perform our selection, we considered all the blazars in the ROMA-BZCAT catalog (Massaro et al. 2010) that are associated with a γ-ray source in the 2FGL (The Fermi-LAT Collaboration 2011). Then, we searched for infrared counterparts in the WISE archive adopting the same criteria described...

Rank	Topic = 20	Topic = 48	Topic = 90
1	entanglement	measured	ferroelectric
2	color	uncertainties	population
3	distance	catalog	rational
4	magnitude	matching	fraction
5	accretion	estimated	starburst
6	similar	respectively	shielding
7	modulus	final	similar
8	objects	cathode	emitting
9	right	total	reputation
10	parameters	limit	respectively

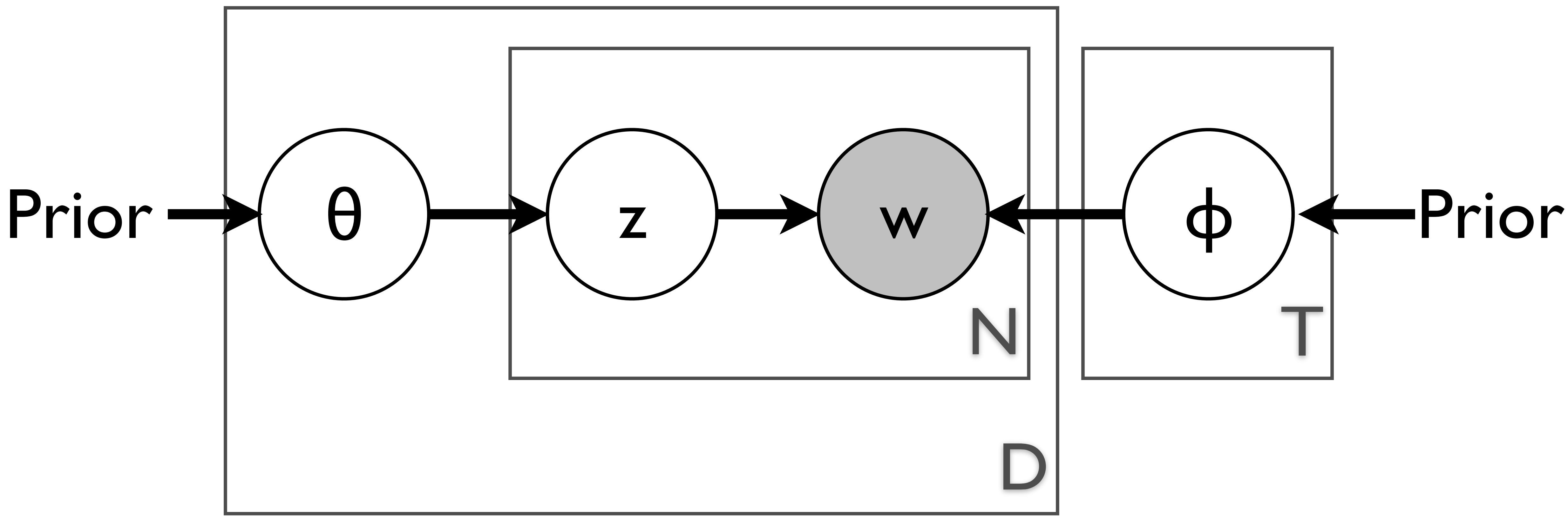
Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)



Modeling Text with Topics

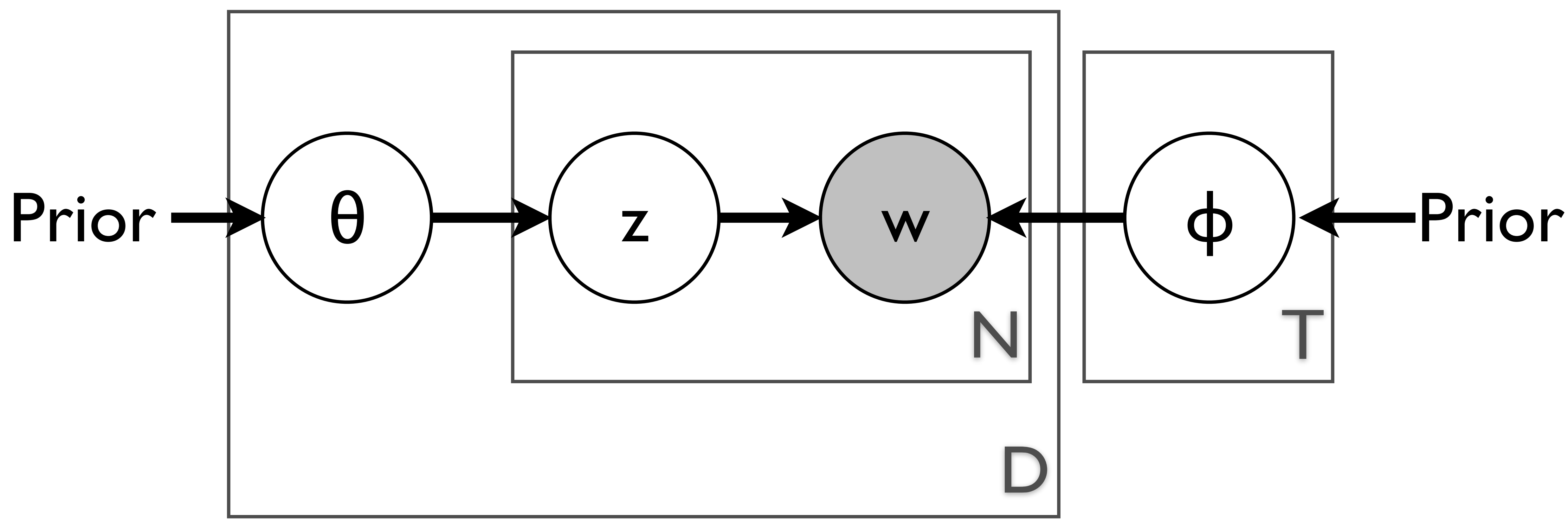
Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)



Multiple languages?

Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

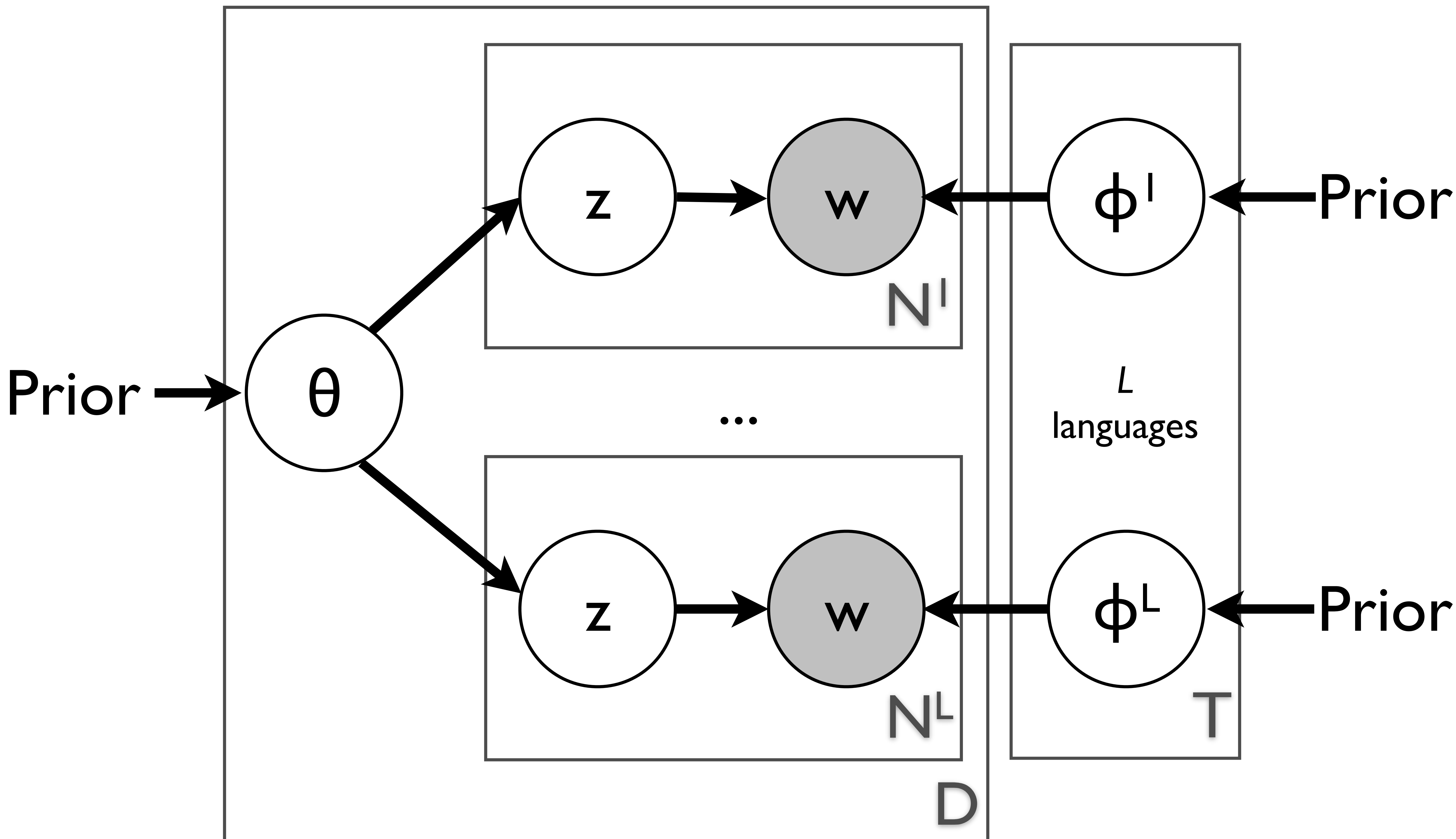


Multiple languages?

graph	problem	rendering	algebra	und	la
graphs	problems	graphics	algebras	von	des
edge	optimization	image	ring	die	le
vertices	algorithm	texture	rings	der	du
edges	programming	scene	modules	im	les

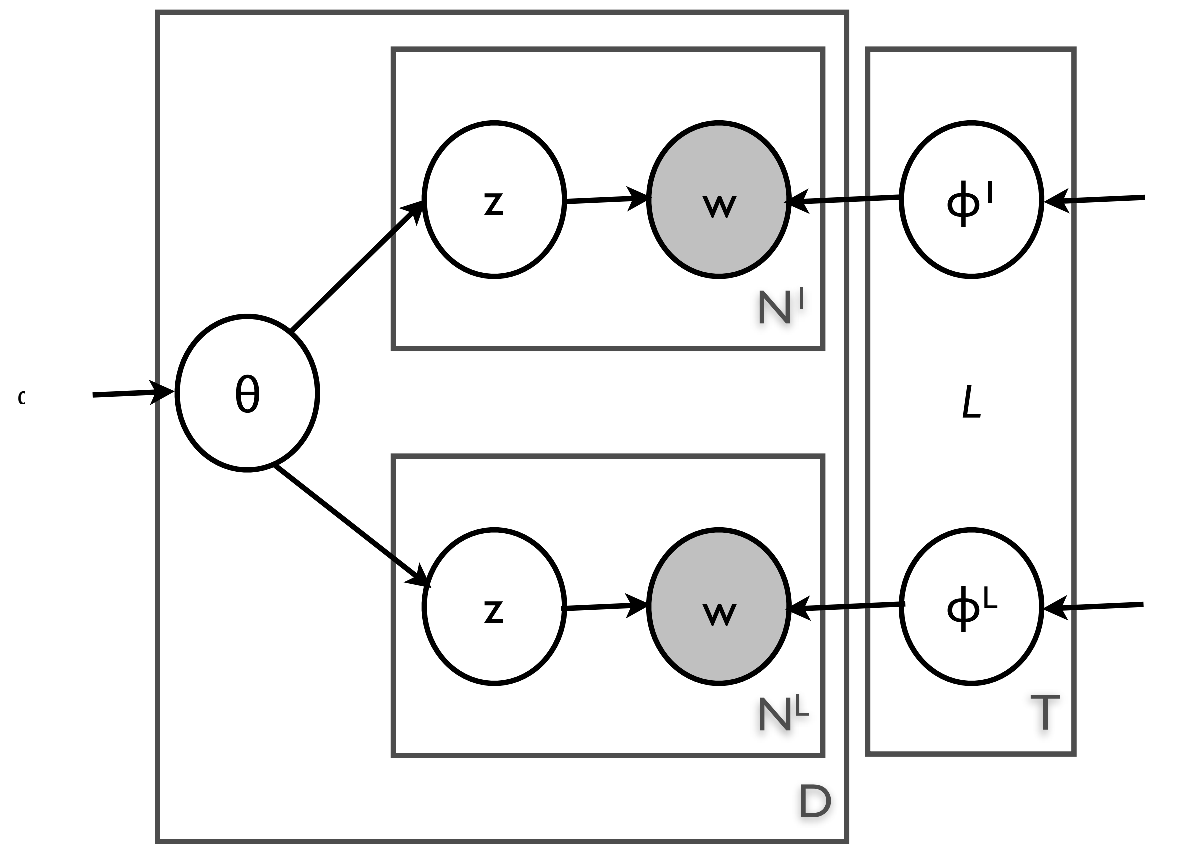
Multilingual Text with Topics

Polylingual Topic Models (EMNLP 2009)



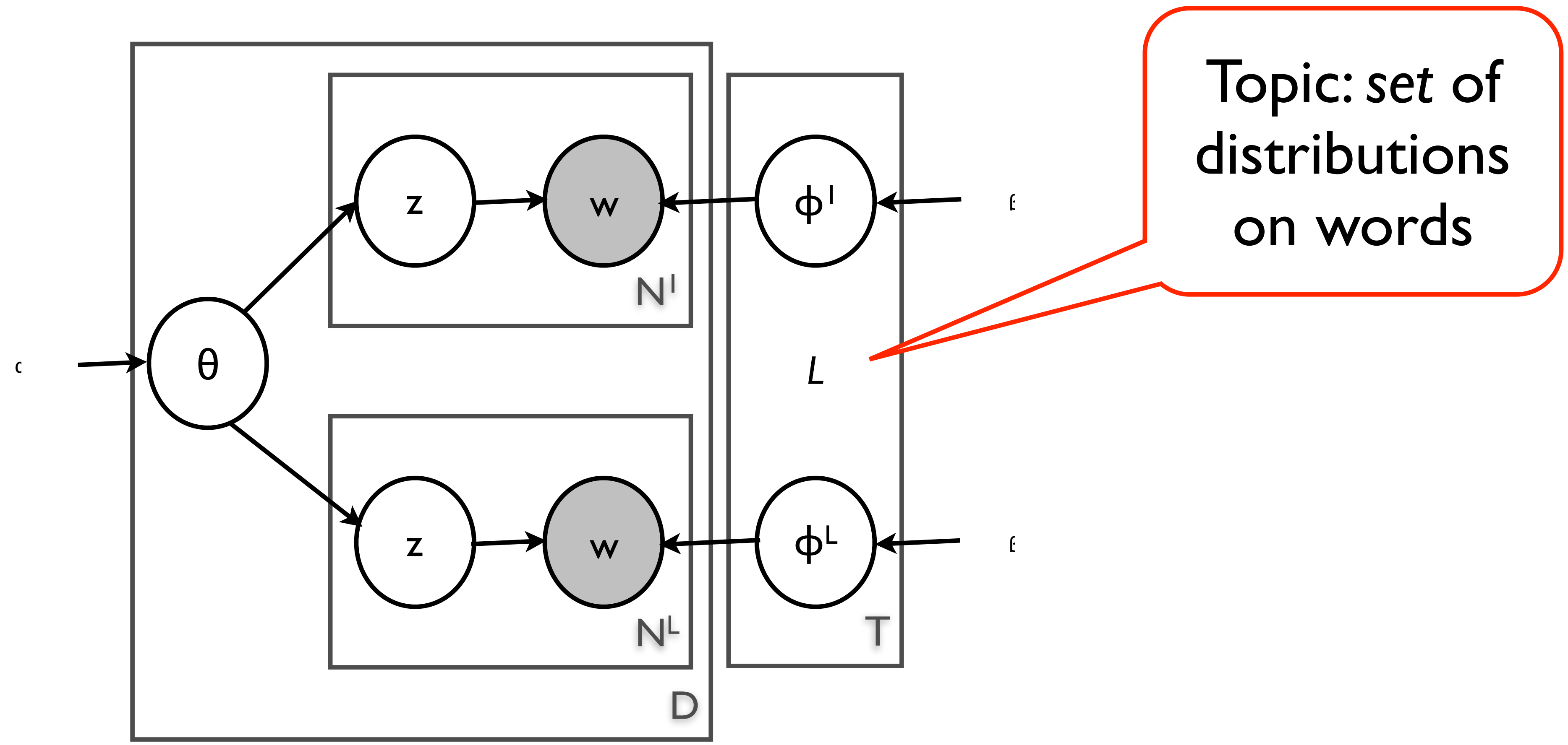
Multilingual Text with Topics

Polylingual Topic Models (EMNLP 2009)



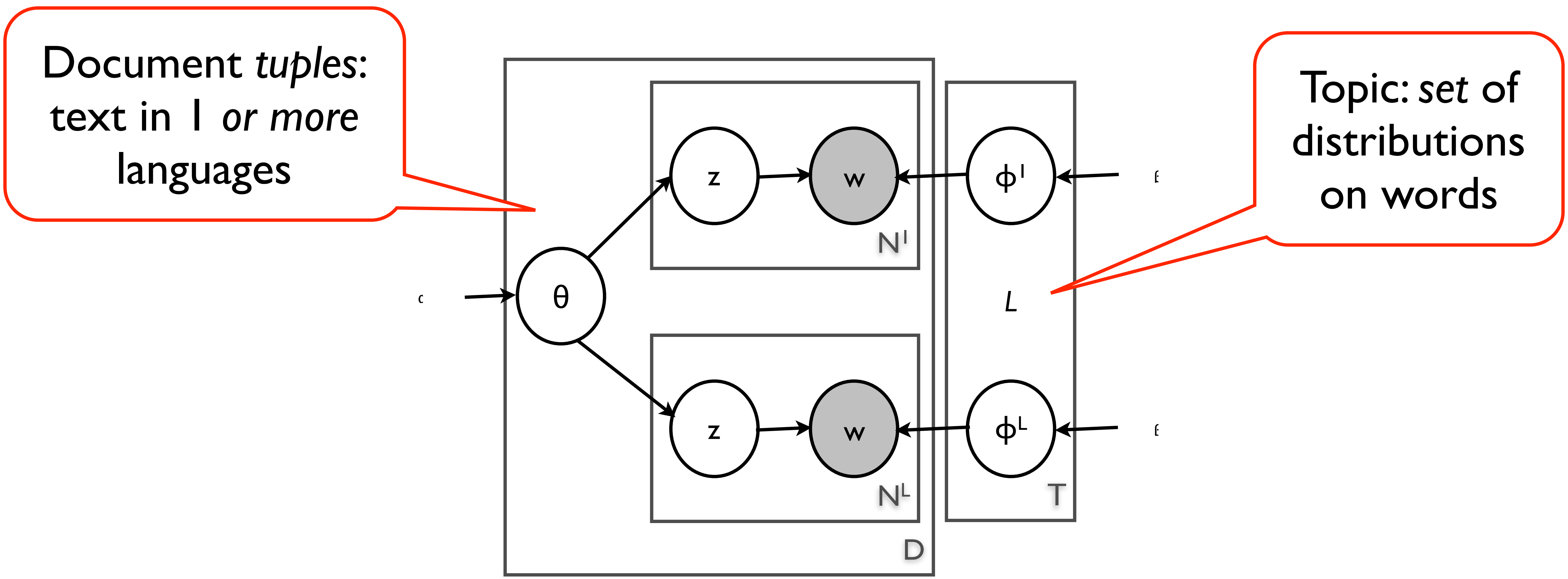
Multilingual Text with Topics

Polylingual Topic Models (EMNLP 2009)



Multilingual Text with Topics

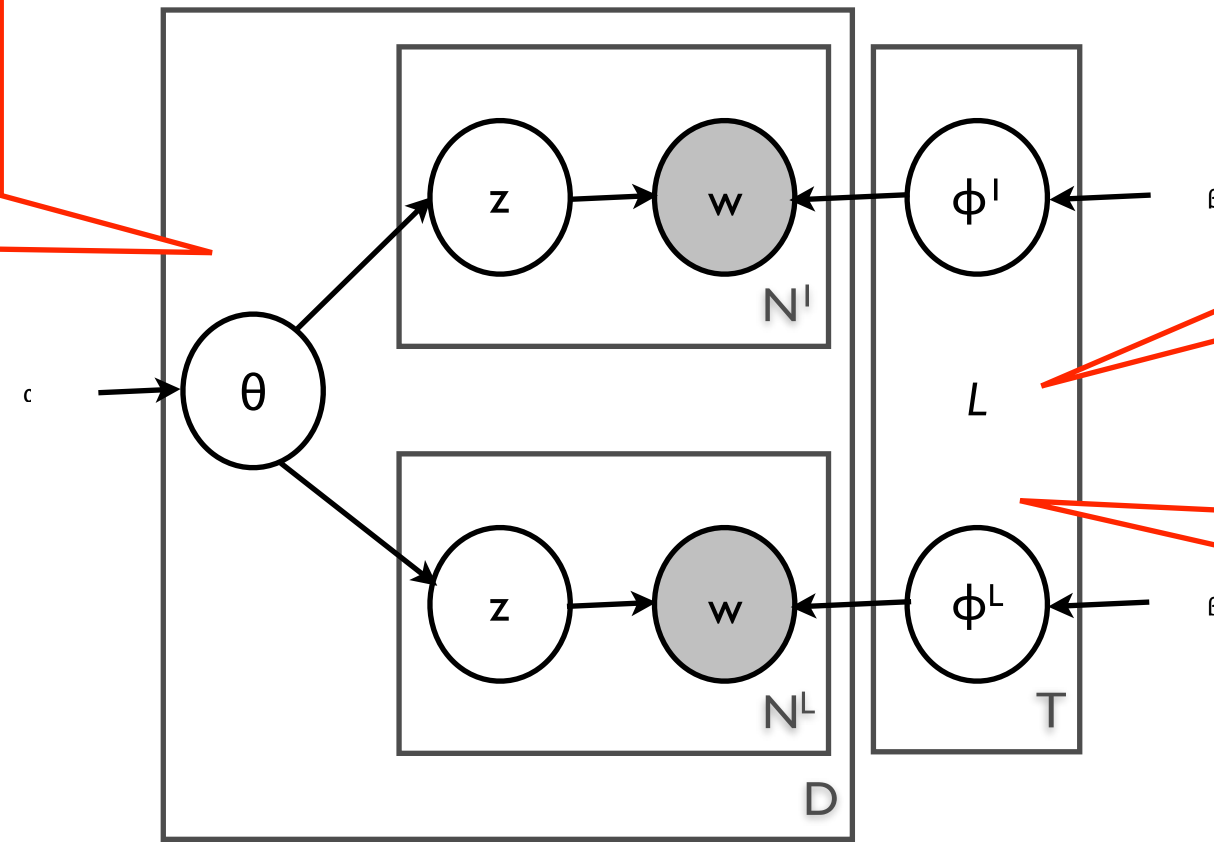
Polylingual Topic Models (EMNLP 2009)



Multilingual Text with Topics

Polylingual Topic Models (EMNLP 2009)

Document *tuples*: text in 1 *or more* languages



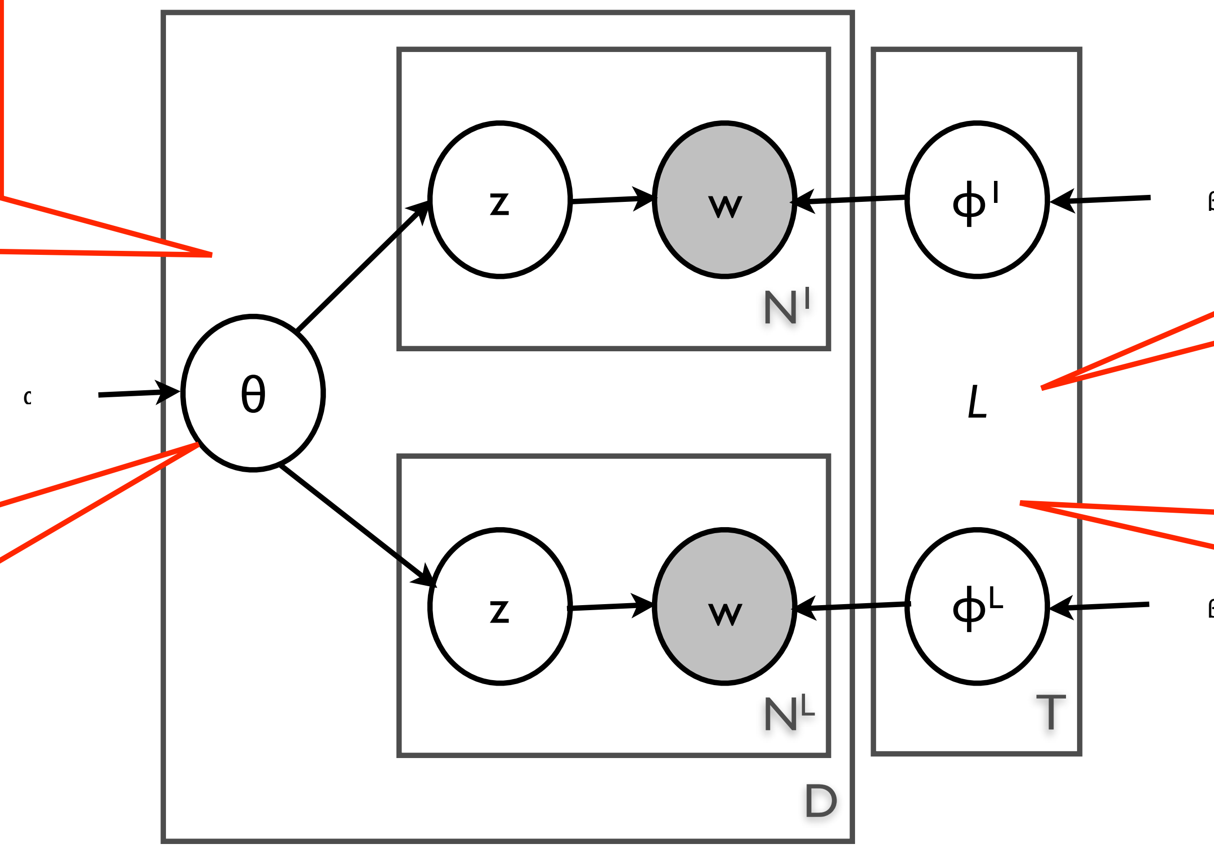
Topic: *set* of distributions on words

Scales *linearly* w/ number of langs. (unlike pairwise)

Multilingual Text with Topics

Polylingual Topic Models (EMNLP 2009)

Document *tuples*: text in 1 *or more* languages



Topic: *set of* distributions on words

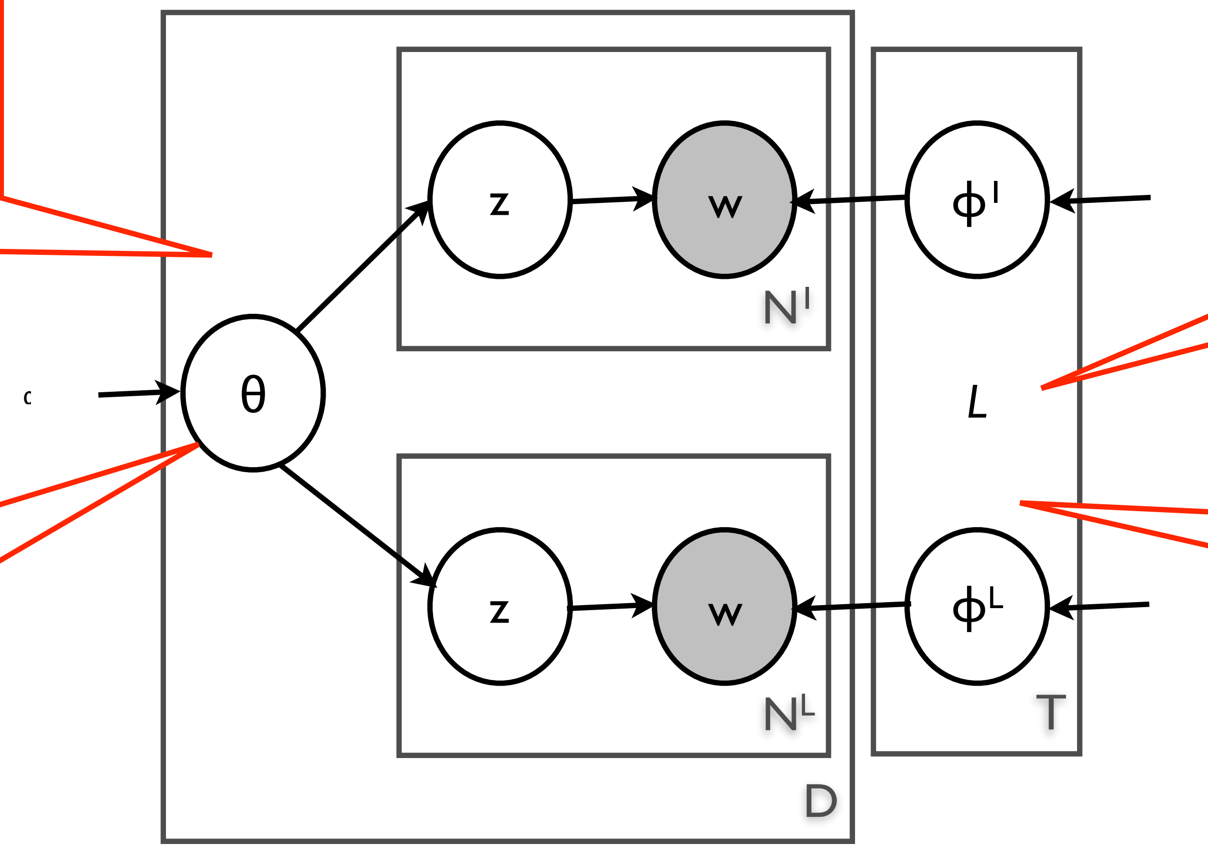
Compare text in different languages with θ, i.e. topic distribution

Scales *linearly* w/ number of langs. (unlike pairwise)

Multilingual Text with Topics

Polylingual Topic Models (EMNLP 2009)

Document *tuples*: text in 1 *or more* languages



Topic: *set of* distributions on words

Compare text in different languages with θ, i.e. topic distribution

Scales *linearly* w/ number of langs. (unlike pairwise)

But...

- No phrase translations
- No distinction of parallel, comparable text
- No modeling of document features (e.g., length)

Parallel Bitext

Genehmigung des Protokolls	Approval of the minutes
Das Protokoll der Sitzung vom Donnerstag, den 28. März 1996 wurde verteilt.	The minutes of the sitting of Thursday, 28 March 1996 have been distributed.
Gibt es Einwände?	Are there any comments?
Die Punkte 3 und 4 widersprechen sich jetzt, obwohl es bei der Abstimmung anders aussah.	Points 3 and 4 now contradict one another whereas the voting showed otherwise.
Das muß ich erst einmal klären, Frau Oomen-Ruijten.	I will have to look into that, Mrs Oomen-Ruijten.

Koehn (2005): European Parliament corpus

Example Europarl Topics

DA	centralbank europæiske ecb s lån centralbanks
DE	zentralbank ezb bank europäischen investitionsbank darlehen
EL	τράπεζα τράπεζας κεντρική εκτ κεντρικής τράπεζες
EN	**bank central ecb banks european monetary**
ES	banco central europeo bce bancos centrales
FI	keskuspankin ekp n euroopan keskuspankki eip
FR	banque centrale bce européenne banques monétaire
IT	banca centrale bce europea banche prestiti
NL	bank centrale ecb europese banken leningen
PT	banco central europeu bce bancos empréstimos
SV	centralbanken europeiska ecb centralbankens s lån

T = 400

Example Europarl Topics

DA	mål nå målsætninger målet målsætning opnå
DE	ziel ziele erreichen zielen erreicht zielsetzungen
EL	στόχους στόχο στόχος στόχων στόχοι επίτευξη
EN	**objective objectives achieve aim ambitious set**
ES	objetivo objetivos alcanzar conseguir lograr estos
FI	tavoite tavoitteet tavoitteena tavoitteiden tavoitteita tavoitteen
FR	objectif objectifs atteindre but cet ambitieux
IT	obiettivo obiettivi raggiungere degli scopo quello
NL	doelstellingen doel doelstelling bereiken bereikt doelen
PT	objectivo objectivos alcançar atingir ambicioso conseguir
SV	mål målet uppnå målen målsättningar målsättning

T = 400

Example Europarl Topics

DA	andre anden side ene andet øvrige
DE	anderen andere einen wie andererseits anderer
EL	άλλες άλλα άλλη άλλων άλλους όπως
EN	**other one hand others another there**
ES	otros otras otro otra parte demás
FI	muiden toisaalta muita muut muihin muun
FR	autres autre part côté ailleurs même
IT	altri altre altro altra dall parte
NL	andere anderzijds anderen ander als kant
PT	outros outras outro lado outra noutros
SV	andra sidan å annat ena annan

T = 400

Multilingual Topical Similarity

Abraham Lincoln

From Wikipedia, the free encyclopedia

This article is about the American president. For other uses, see Abraham Lincoln (disambiguation).

Abraham Lincoln /ˈeɪbrəhæm ˈlɪŋkən/ (February 12, 1809 – April 15, 1865) was the 16th President of the United States, serving from March 1861 until his assassination in April 1865. He successfully led his country through a great constitutional, military and moral crisis – the American Civil War – preserving the Union, while ending slavery, and promoting economic and financial modernization. Reared in a poor family on the western frontier, Lincoln was mostly self-educated. He became a country lawyer, an Illinois state legislator, and a one-term member of the United States House of Representatives, but failed in two attempts to be elected to the United States Senate.

Abraham Lincoln

Abraham Lincoln [ˈeɪbrəhæm ˈliŋkən] (* 12. Februar 1809 bei Hodgenville, Hardin County, heute: LaRue County, Kentucky; † 15. April 1865 in Washington, D.C.) amtierte von 1861 bis 1865 als 16. Präsident der Vereinigten Staaten von Amerika. Er war der erste aus den Reihen der Republikanischen Partei und der erste, der einem Attentat zum Opfer fiel. 1860 gewählt, gelang ihm 1864 die Wiederwahl.

Seine Präsidentschaft gilt als eine der bedeutendsten in der Geschichte der Vereinigten Staaten: Die Wahl des Sklavereigegners veranlasste zunächst sieben, später weitere vier der sklavenhaltenden Südstaaten zur Sezession. Lincoln führte die verbliebenen Nordstaaten durch den daraus entstandenen Bürgerkrieg, setzte die Wiederherstellung der Union durch und betrieb erfolgreich die Abschaffung der Sklaverei in den Vereinigten Staaten. Unter seiner Regierung schlugen die USA den Weg zum zentral regierten, modernen Industriestaat ein und schufen so die Basis für ihren Aufstieg zur Weltmacht im 20. Jahrhundert.

Example Wikipedia Topics

CY	sadwrn blaned gallair at lloeren mytholeg
DE	space nasa sojus flug mission
EL	διαστημικό sts nasa αγγλ small
EN	**space mission launch satellite nasa spacecraft**
FA	فضایی ماموریت ناسا مدار فضانورد ماهواره
FI	sojuz nasa apollo ensimmäinen space lento
FR	spatiale mission orbite mars satellite spatial
HE	החלל הארץ חלל כדור א תוכנית
IT	spaziale missione programma space sojuz stazione
PL	misja kosmicznej stacji misji space nasa
RU	космический союз космического спутник станции
TR	uzay soyuz ay uzaya salyut sovyetler

$T = 400$

Example Wikipedia Topics

CY	sbaen madrid el la josé sbaeneg
DE	de spanischer spanischen spanien madrid la
EL	ισπανίας ισπανία de ισπανός ντε μαδρίτη
EN	**de spanish spain la madrid y**
FA	ترین de اسپانیا اسپانیایی کوبا مادرید
FI	espanja de espanjan madrid la real
FR	espagnol espagne madrid espagnole juan y
HE	ספרד ספרדית דה מדריד הספרדית קובה
IT	de spagna spagnolo spagnola madrid el
PL	de hiszpański hiszpanii la juan y
RU	де мадрид испании испания испанский de
TR	ispanya ispanyol madrid la küba real

T = 400

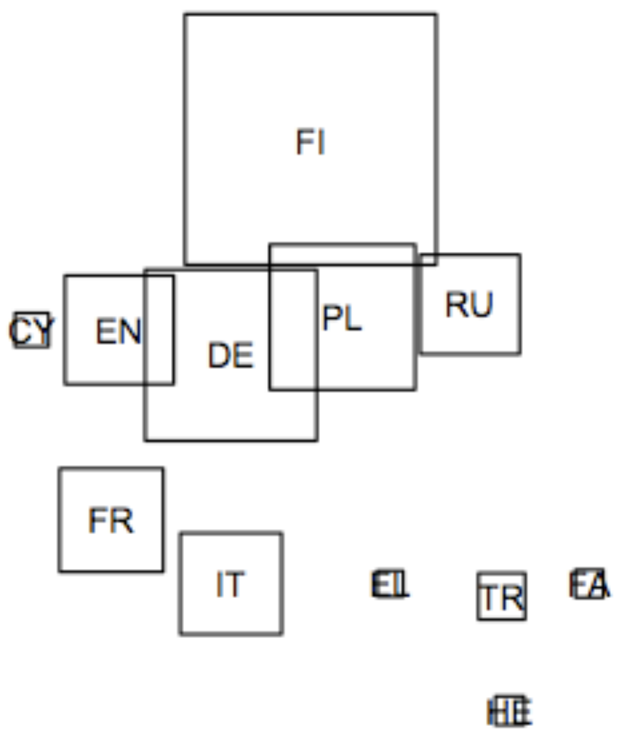
Example Wikipedia Topics

CY	bardd gerddi iaith beirdd fardd gymraeg
DE	dichter schriftsteller literatur gedichte gedicht werk
EL	ποιητής ποίηση ποιητή έργο ποιητές ποιήματα
EN	**poet poetry literature literary poems poem**
FA	شاعر شعر ادبیات فارسی ادبی آثار
FI	runoilija kirjailija kirjallisuuden kirjoitti runo julkaisi
FR	poète écrivain littérature poésie littéraire ses
HE	משורר ספרות שירה סופר שירים המשורר
IT	poeta letteratura poesia opere versi poema
PL	poeta literatury poezji pisarz in jego
RU	поэт его писатель литературы поэзии драматург
TR	şair edebiyat şiir yazar edebiyatı adlı

T = 400

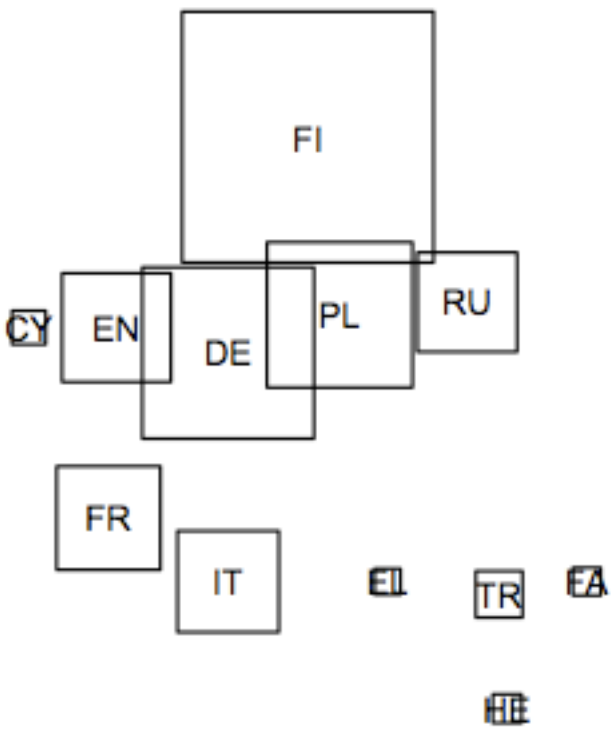
Differences in Topic Emphasis

Differences in Topic Emphasis

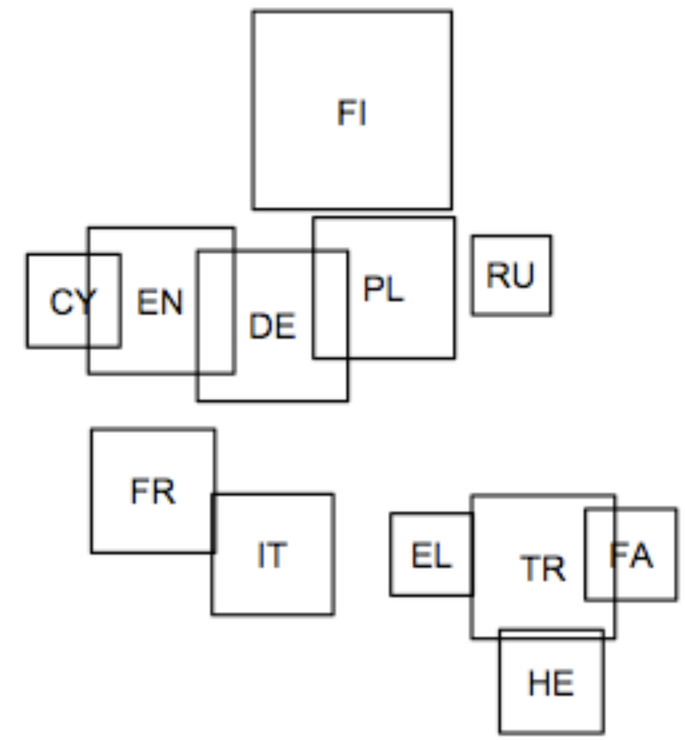


world ski km won

Differences in Topic Emphasis

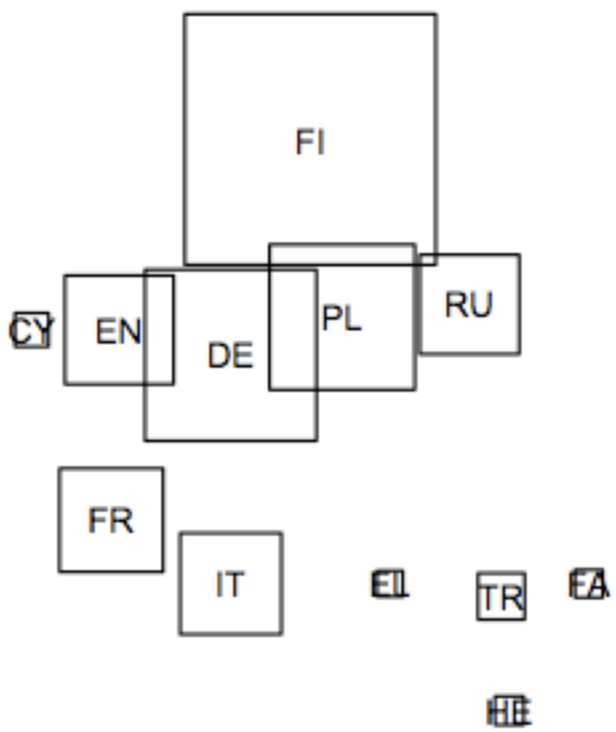


world ski km won

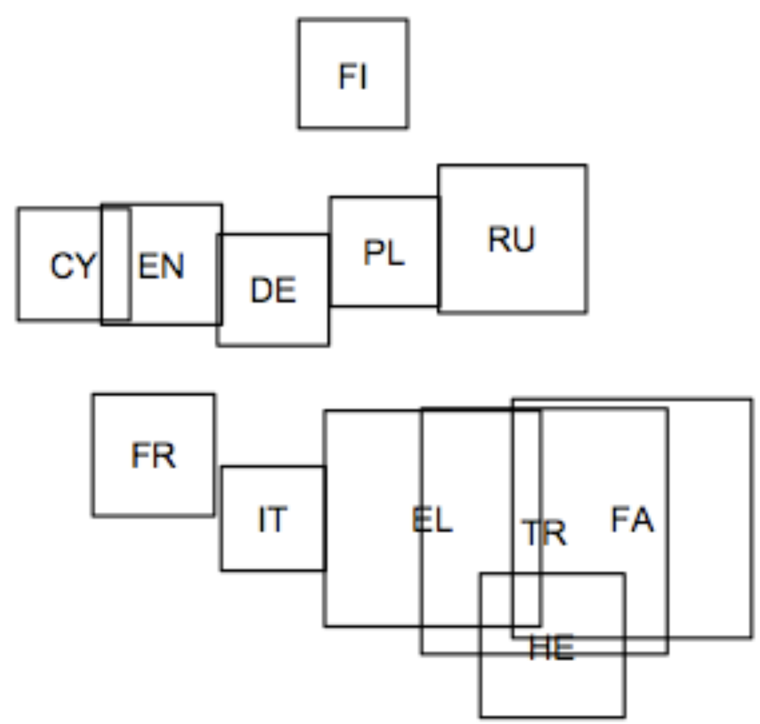


actor role television actress

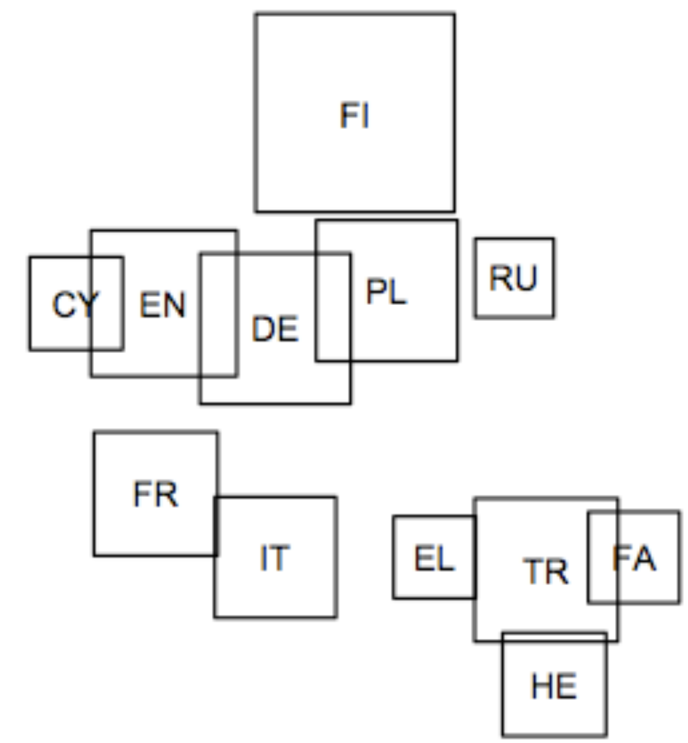
Differences in Topic Emphasis



world ski km won



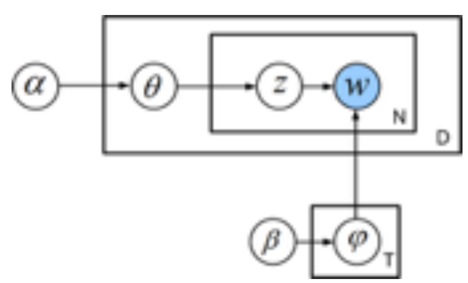
ottoman empire khan byzantine



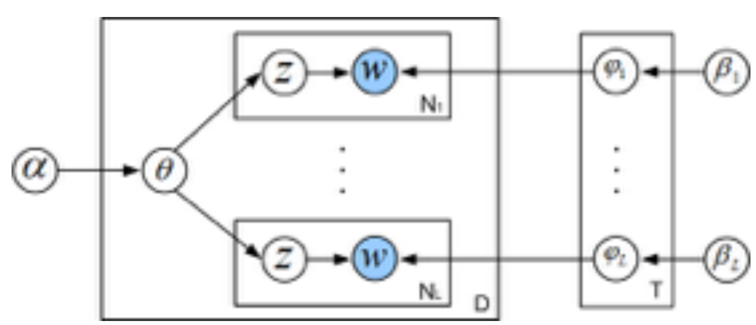
actor role television actress

Document Inference

Latent Dirichlet Allocation (LDA)

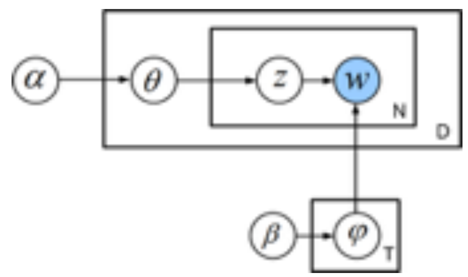


Polylingual Topic Model (PLTM)

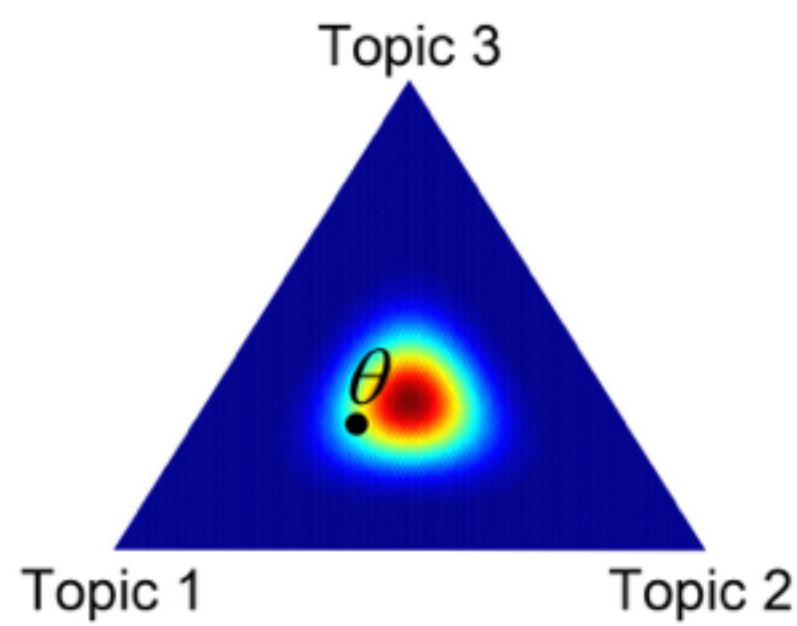


Document Inference

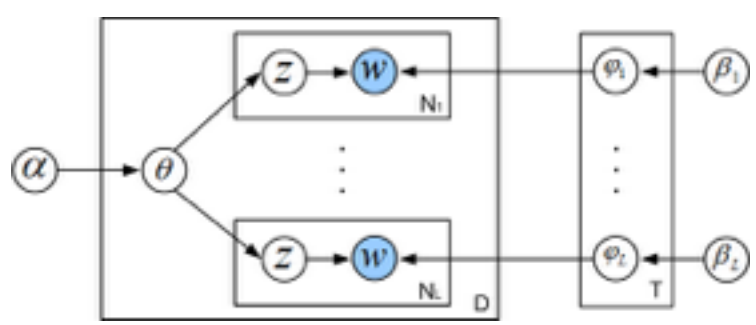
Latent Dirichlet Allocation (LDA)



Inference



Polylingual Topic Model (PLTM)



Bootstrapping Translation Detection and Sentence Extraction

- Extracted English-Spanish news stories from the Gigaword collection using PLTM trained on OCD output:

EN: WASHINGTON, URGENT: Treasury chief defends dollar as world reserve currency. US Treasury Secretary Timothy Geithner said Wednesday that "the dollar remains the world's standard reserve currency", following China's call for a new global currency as an alternative to the greenback.

He(EN,ES)=0.055

ES: WASHINGTON, URGENTE: Washington quiere que el dólar se mantenga como la principal divisa de reserve. El secretario del Tesoro estadou-nidense Timothy Geithner declaró este miércoles que el dólar se mantiene como la principal moneda mundial de reserva y que Estados Unidos bregará porque se mantenga como tal.

He(EN,ES)=0.086

ES: Washington: EEUU quiere que el dólar se mantenga como la principal divisa de reserva. El secretario del Tesoro estadounidense Timothy Geithner declaró este miércoles que el dólar se mantiene como la principal moneda mundial de reserva y que Estados Unidos bregará porque se mantenga como tal. "Pienso que el dólar sigue siendo la moneda de reserva de referencia y pienso que debería continuar siéndolo durante largo tiempo", declaró Geithner ante el Consejo de Relaciones Exteriores en Nueva York. "Como país haremos lo necesario para conservar la confianza en nuestros mercados financieros" y en nuestra economía, agregó.

He(EN,ES)=0.153

ES: BUENOS AIRES: Peso argentino estable a 3,70 por dólar. La moneda argentina se mantuvo estable este miércoles a 3,70 pesos por dólar, según el promedio de bancos y casas de cambio. El Banco Central viene interviniendo en el mercado para administrar una devaluación gradual de la moneda con respecto al dólar estadounidense.

He(EN,ES)=0.172

ES: WASHINGTON: Obama defiende derecho a la expansión de la OTAN. El presidente estadou-nidense Barack Obama dijo este miércoles que Estados Unidos quería "reiniciar" las relaciones con Rusia pero añadió que la OTAN debería de todos modos estar abierta a los países que aspiren a unirse a esa alianza. "Mi gobierno busca reiniciar las relaciones con Rusia", dijo Obama al cabo de una reunión en la Casa Blanca con el secretario general de la OTAN, Jaap de Hoop Scheffer. Pero dijo que los renovados vínculos con Moscú deben ser "consistentes con la membresía de la OTAN y consistentes con la necesidad de enviar una clara señal en Europa de que vamos a atenernos (...)

Training MT from Comparable Corpora

- MT system performance - parallel vs. extracted sentences
  - Parallel collection: News Commentary(all) & Europarl(all)
  - Extracted Sentences: Gigaword (4 years)

Training Source	Collection Size		Test Set	
	Parallel	Extracted	News (WMT'11)	Europarl (WMT'09)
News Commentary (NC)	131K	0	23.75	25.43
Europarl (EU)	1.75M	0	23.91	32.06
Gigaword Extracted (GE)	0	926K	**24.25**	23.88
NC+GE	131K	926K	24.92	25.61
EU+GE	1.75M	926K	25.90	31.59

Krstovski, 2016

Bilingual Embeddings

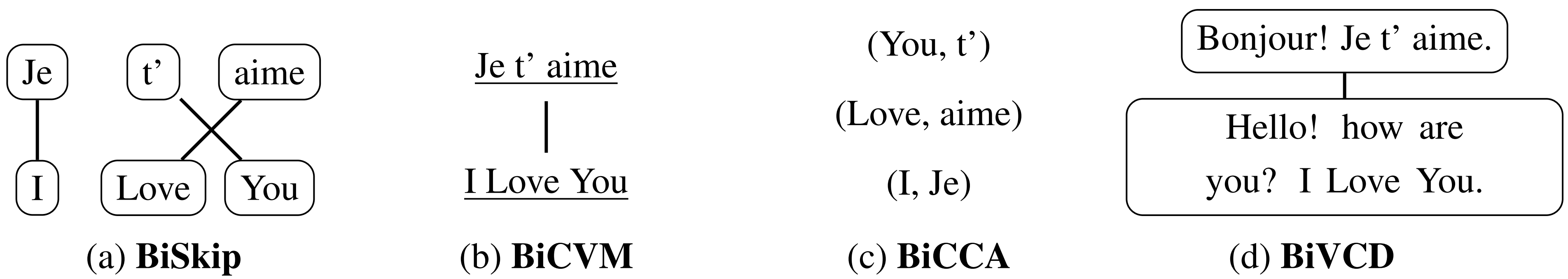
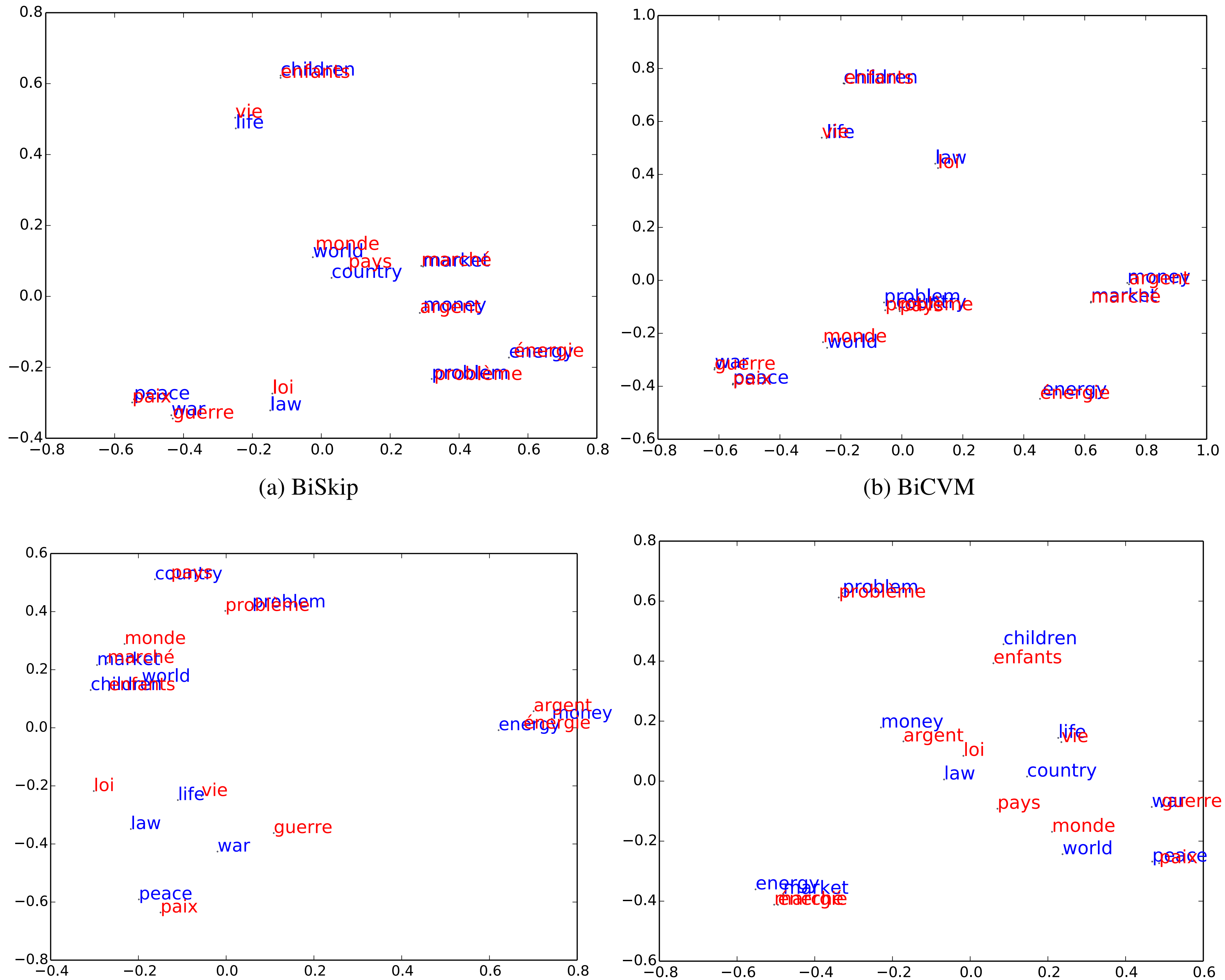


Figure 2: Forms of supervision required by the four models compared in this paper. From left to right, the cost of the supervision required varies from expensive (BiSkip) to cheap (BiVCD). BiSkip requires a parallel corpus annotated with word alignments (Fig. 2a), BiCVM requires a sentence-aligned corpus (Fig. 2b), BiCCA only requires a bilingual lexicon (Fig. 2c) and BiVCD requires comparable documents (Fig. 2d).

Upadhyay et al. (2016)

Bilingual Embeddings



(a) BiSkip (b) BiCVM

(c) BiCCA (d) BiVCD

Upadhyay et al. (2016)

Search

What’s the best translation
(under our model)?

Search

- Even if we know the right words in a translation, there are *n*! permutations.

$10! = 3,626,800$ $20! \approx 2.43 \times 10^{18}$ $30! \approx 2.65 \times 10^{32}$

- We want the translation that gets the highest score under our model
  - Or the best *k* translations
  - Or a random sample from the model's distribution
- But **not** in *n*! time!

Search in Phrase Models

Deshalb haben wir allen Grund , die Umwelt in die Agrarpolitik zu integrieren

Translate in target language order to ease language modeling.

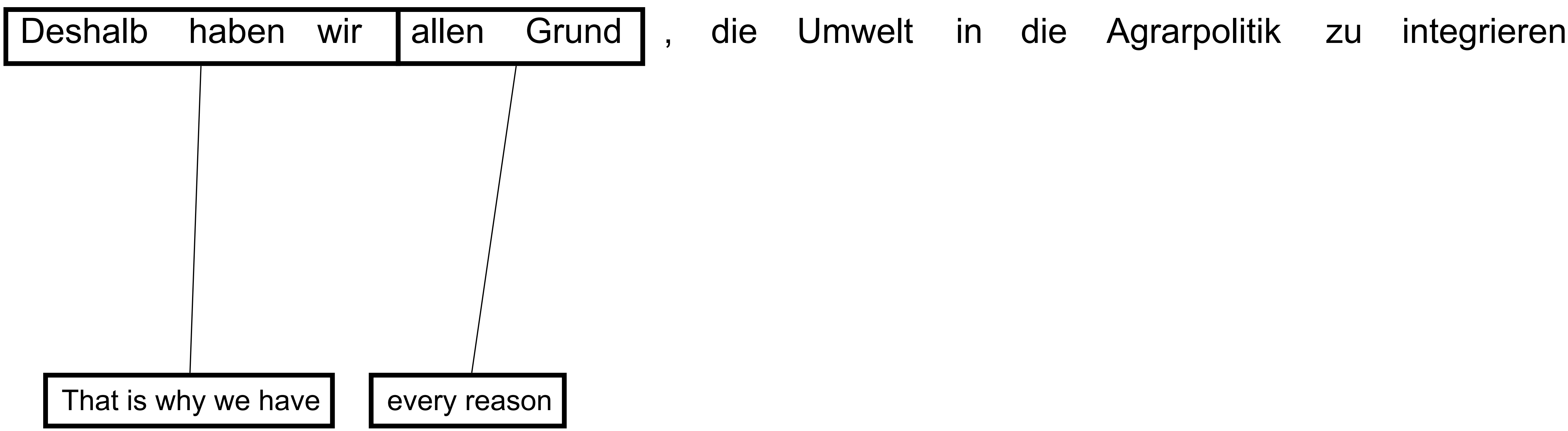
Search in Phrase Models

Deshalb haben wir allen Grund , die Umwelt in die Agrarpolitik zu integrieren

That is why we have

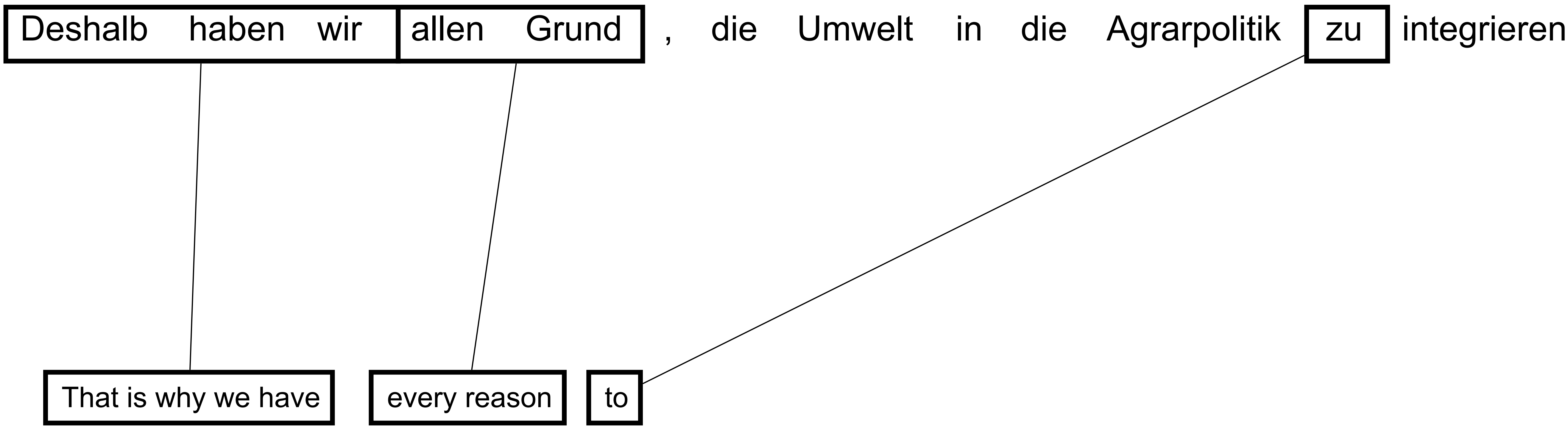
Translate in target language order to ease language modeling.

Search in Phrase Models



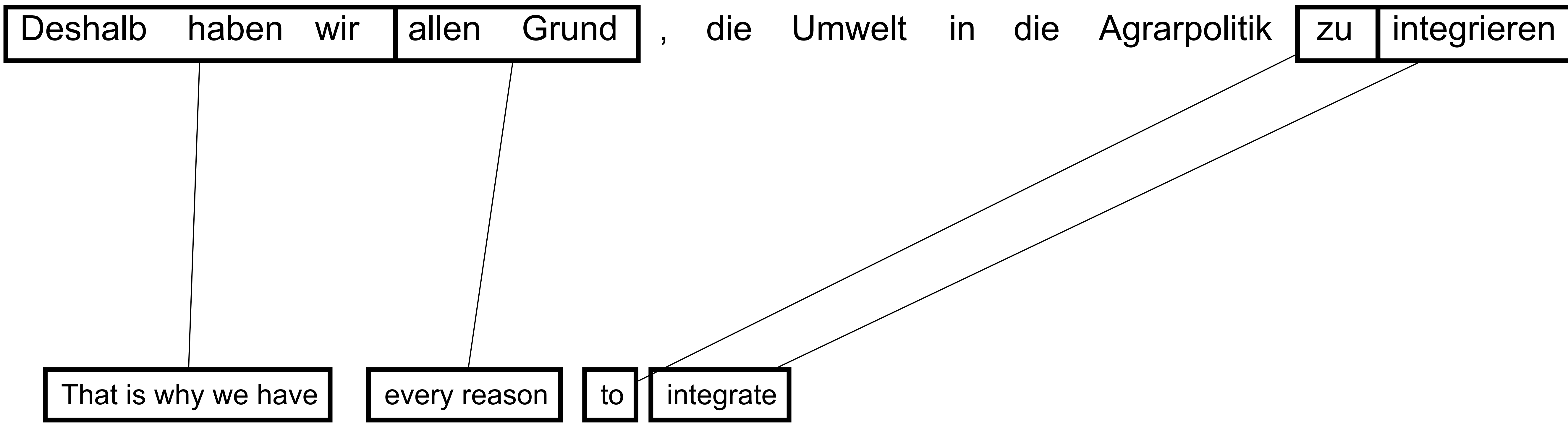
Translate in target language order to ease language modeling.

Search in Phrase Models



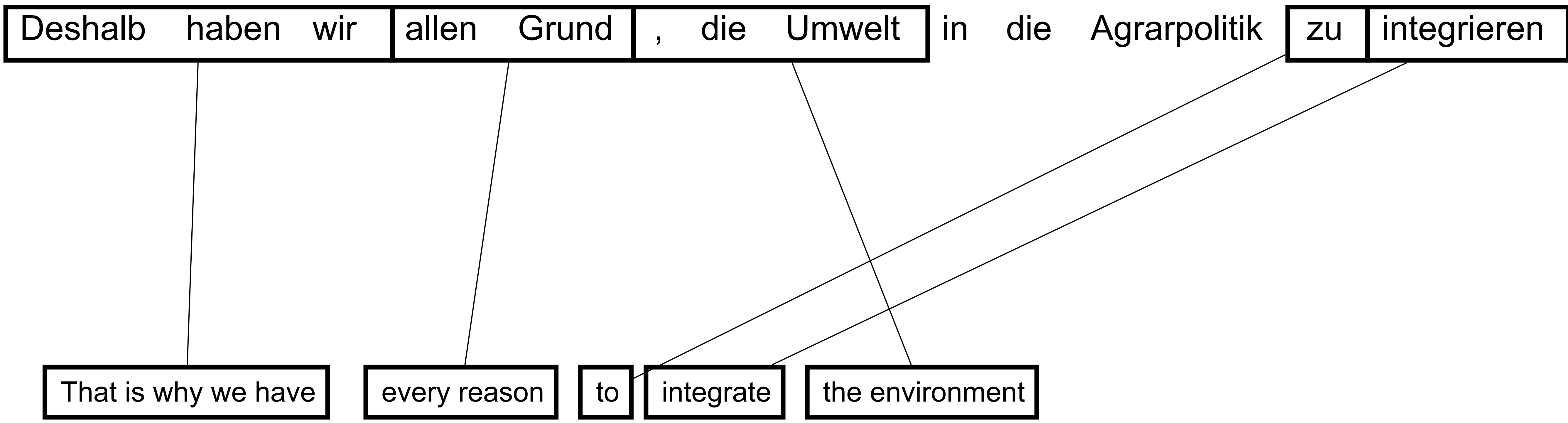
Translate in target language order to ease language modeling.

Search in Phrase Models



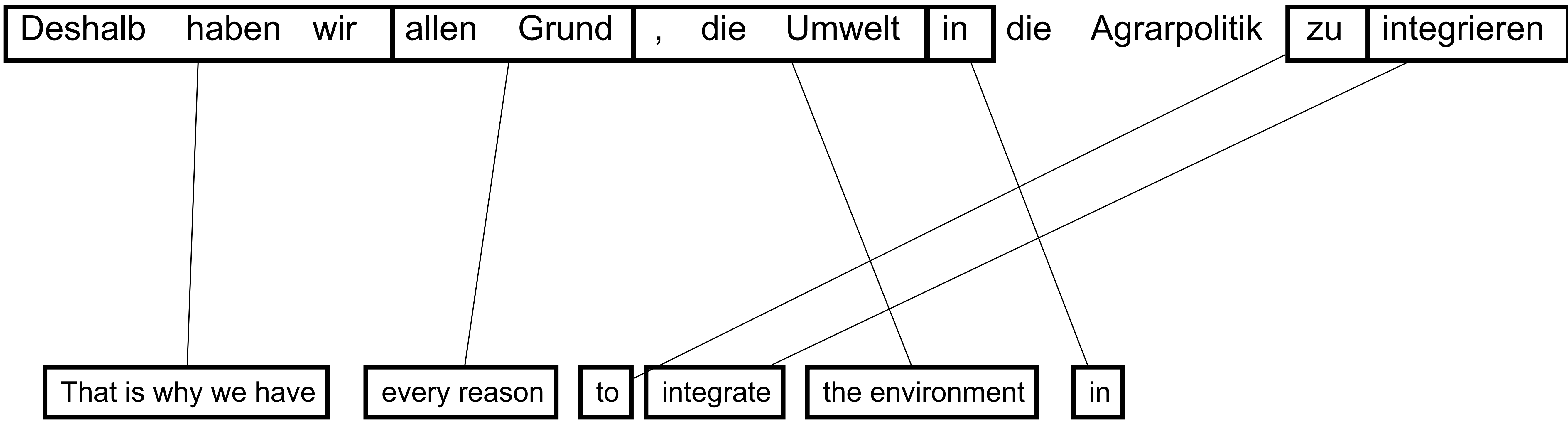
Translate in target language order to ease language modeling.

Search in Phrase Models



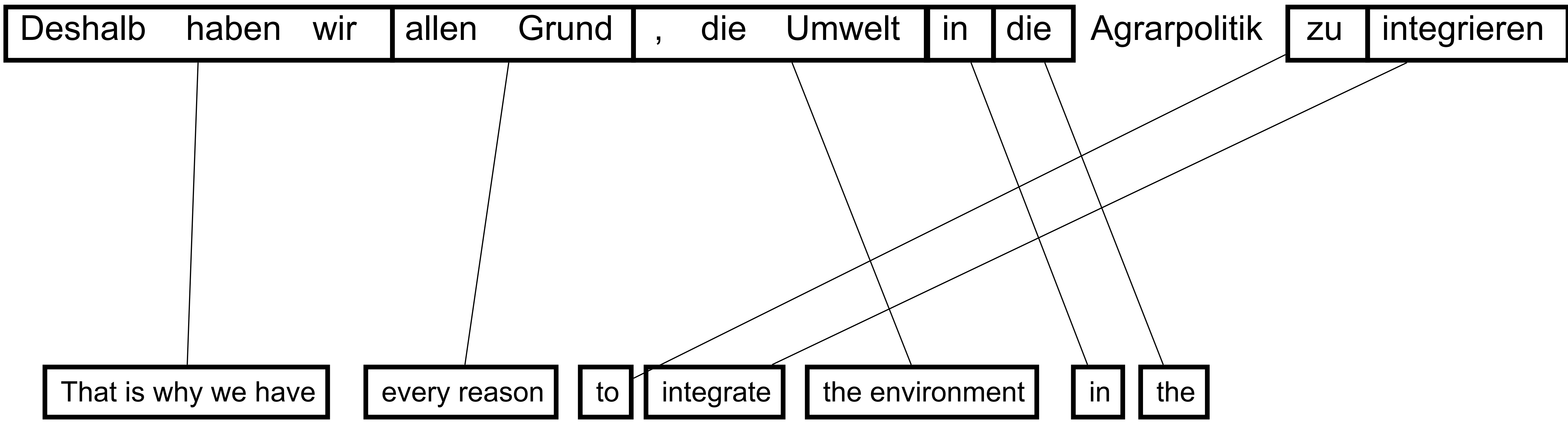
Translate in target language order to ease language modeling.

Search in Phrase Models



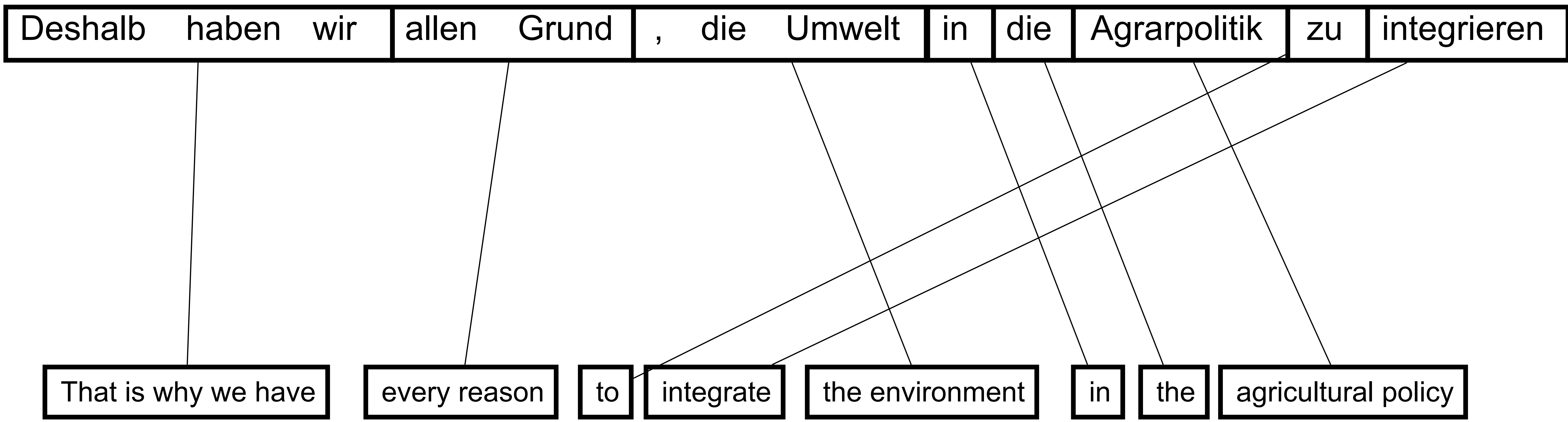
Translate in target language order to ease language modeling.

Search in Phrase Models



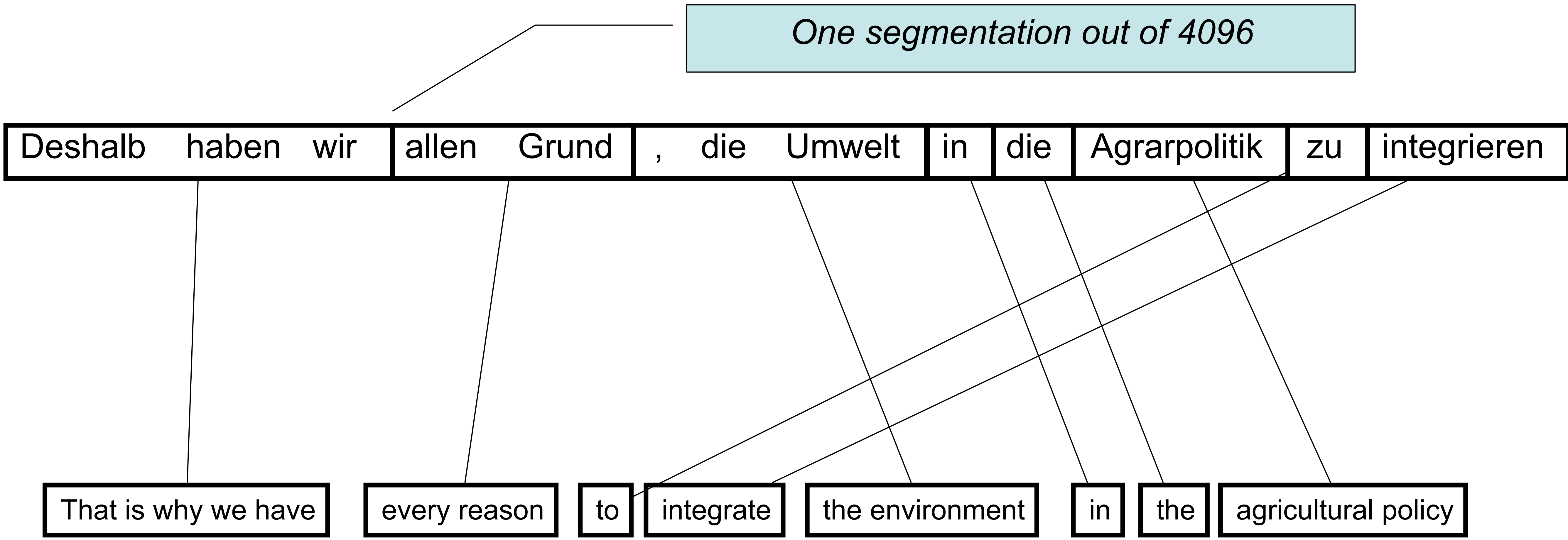
Translate in target language order to ease language modeling.

Search in Phrase Models



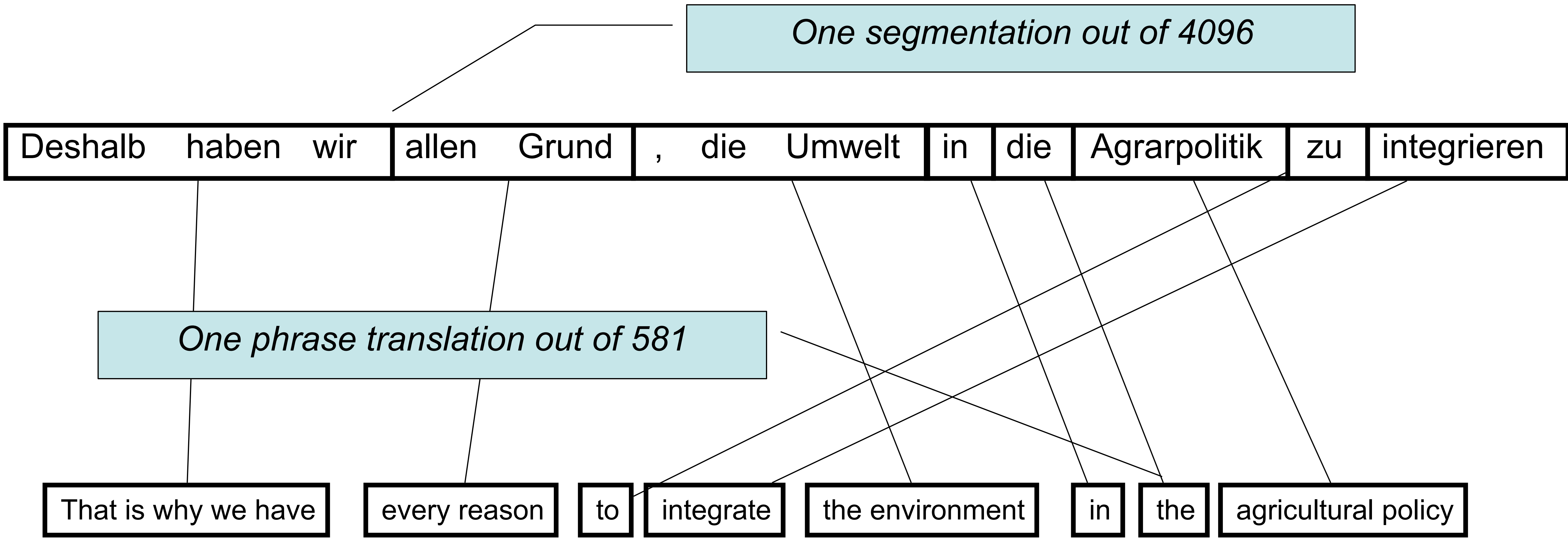
Translate in target language order to ease language modeling.

Search in Phrase Models



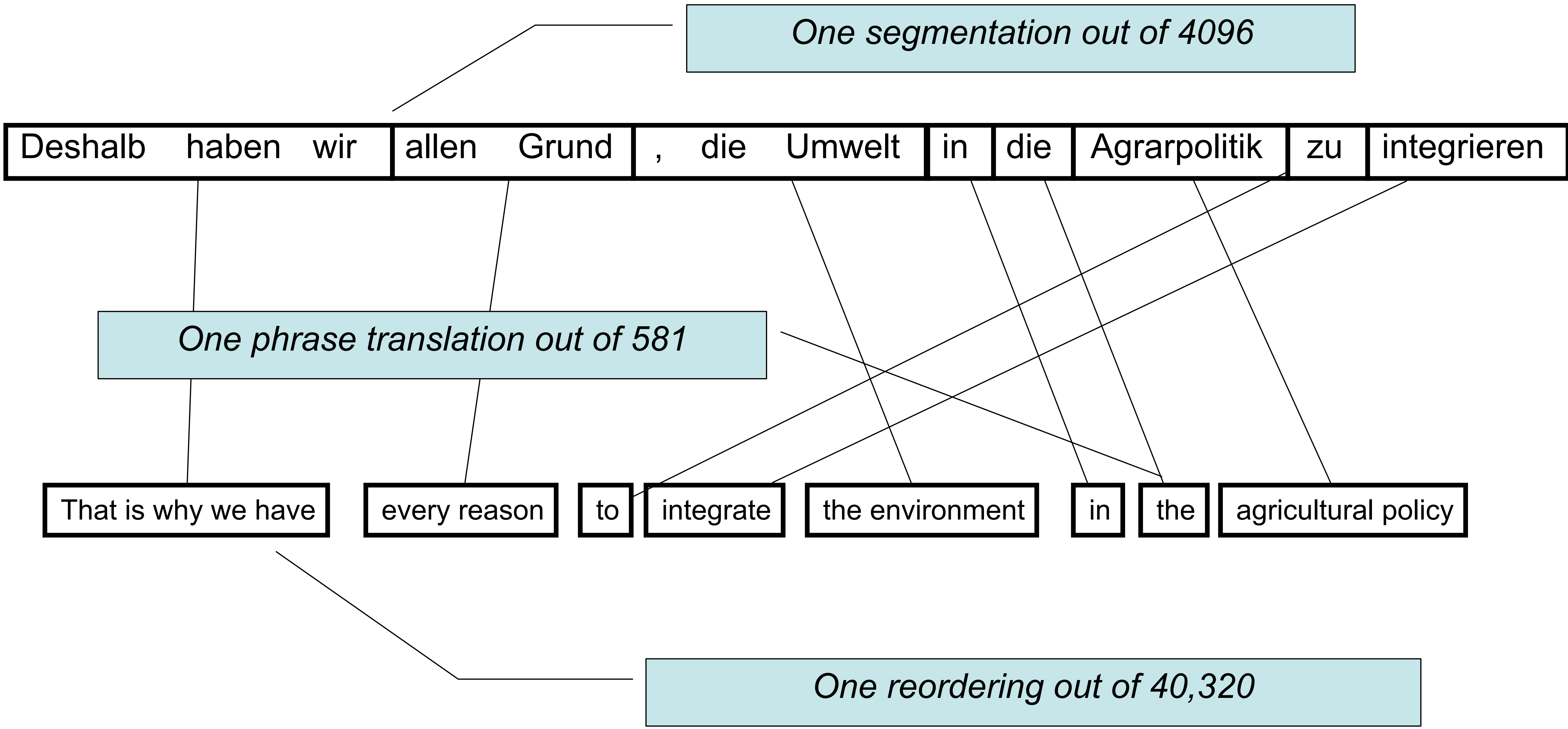
Translate in target language order to ease language modeling.

Search in Phrase Models



Translate in target language order to ease language modeling.

Search in Phrase Models



Translate in target language order to ease language modeling.

Search in Phrase Models

<table>
<tr><th>Deshalb</th><th>haben</th><th>wir</th><th>allen</th><th>Grund</th><th>,</th><th>die</th><th>Umwelt</th><th>in</th><th>die</th><th>Agrarpolitik</th><th>zu</th><th>integrieren</th></tr>
<tr><td colspan="3">that is why we have</td><td colspan="2">every reason</td><td colspan="3">the environment</td><td>in</td><td>the</td><td>agricultural policy</td><td>to</td><td>integrate</td></tr>
<tr><td>therefore</td><td>have</td><td>we</td><td colspan="3">every reason</td><td>the</td><td>environment</td><td colspan="2">in the</td><td>agricultural policy ,</td><td colspan="2">to integrate</td></tr>
<tr><td>that is why</td><td colspan="2">we have</td><td>all</td><td>reason</td><td>,</td><td>which</td><td colspan="2">environment in</td><td colspan="2">agricultural policy</td><td colspan="2">parliament</td></tr>
<tr><td colspan="2">have therefore</td><td>us</td><td>all the</td><td colspan="2">reason</td><td>of the</td><td colspan="2">environment into</td><td colspan="2">the agricultural policy</td><td colspan="2">successfully integrated</td></tr>
<tr><td colspan="2">hence</td><td>, we</td><td>every</td><td colspan="3">reason to make</td><td>environmental</td><td>on</td><td colspan="2">the cap</td><td colspan="2">be woven together</td></tr>
<tr><td colspan="3">we have therefore</td><td>everyone</td><td colspan="3">grounds for taking the</td><td>the environment</td><td colspan="2">to the</td><td>agricultural policy is</td><td>on</td><td>parliament</td></tr>
<tr><td>so</td><td colspan="2">, we</td><td>all of</td><td>cause</td><td colspan="2">which</td><td>environment ,</td><td>to</td><td colspan="2">the cap ,</td><td>for</td><td>incorporated</td></tr>
<tr><td colspan="3">hence our</td><td>any</td><td>why</td><td colspan="2">that</td><td>outside</td><td>at</td><td colspan="2">agricultural policy</td><td>too</td><td>woven together</td></tr>
<tr><td colspan="2">therefore ,</td><td>it</td><td>of all</td><td>reason for</td><td colspan="2">, the</td><td>completion</td><td>into</td><td colspan="2">that agricultural policy</td><td>be</td><td></td></tr>
</table>

And many, many more…even before reordering

Search in Phrase Models

<table>
<tr><th>Deshalb</th><th>haben</th><th>wir</th><th>allen</th><th>Grund</th><th>,</th><th>die</th><th>Umwelt</th><th>in</th><th>die</th><th>Agrarpolitik</th><th>zu</th><th>integrieren</th></tr>
<tr><td colspan="3">that is why we have</td><td colspan="2">every reason</td><td colspan="3">the environment</td><td>in</td><td>the</td><td>agricultural policy</td><td>to</td><td>integrate</td></tr>
<tr><td>therefore</td><td>have</td><td>we</td><td colspan="3">every reason</td><td>the</td><td>environment</td><td colspan="2">in the</td><td>agricultural policy ,</td><td colspan="2">to integrate</td></tr>
<tr><td>that is why</td><td colspan="2">we have</td><td>all</td><td>reason</td><td>,</td><td>which</td><td colspan="2">environment in</td><td colspan="2">agricultural policy</td><td colspan="2">parliament</td></tr>
<tr><td colspan="2">have therefore</td><td>us</td><td>all the</td><td colspan="2">reason</td><td>of the</td><td colspan="2">environment into</td><td colspan="2">the agricultural policy</td><td colspan="2">successfully integrated</td></tr>
<tr><td colspan="2">hence</td><td>, we</td><td>every</td><td colspan="3">reason to make</td><td>environmental</td><td>on</td><td colspan="2">the cap</td><td colspan="2">be woven together</td></tr>
<tr><td colspan="3">we have therefore</td><td>everyone</td><td colspan="3">grounds for taking the</td><td>the environment</td><td colspan="2">to the</td><td>agricultural policy is</td><td>on</td><td>parliament</td></tr>
<tr><td>so</td><td colspan="2">, we</td><td>all of</td><td>cause</td><td colspan="2">which</td><td>environment ,</td><td>to</td><td colspan="2">the cap ,</td><td>for</td><td>incorporated</td></tr>
<tr><td colspan="3">hence our</td><td>any</td><td>why</td><td colspan="2">that</td><td>outside</td><td>at</td><td colspan="2">agricultural policy</td><td>too</td><td>woven together</td></tr>
<tr><td colspan="2">therefore ,</td><td>it</td><td>of all</td><td>reason for</td><td colspan="2">, the</td><td>completion</td><td>into</td><td colspan="2">that agricultural policy</td><td>be</td><td></td></tr>
</table>

And many, many more…even before reordering

Search in Phrase Models

<table>
<tr><th>Deshalb</th><th>haben</th><th>wir</th><th>allen</th><th>Grund</th><th>,</th><th>die</th><th>Umwelt</th><th>in</th><th>die</th><th>Agrarpolitik</th><th>zu</th><th>integrieren</th></tr>
<tr><td colspan="3">that is why we have</td><td colspan="2">every reason</td><td colspan="3">the environment</td><td>in</td><td>the</td><td>agricultural policy</td><td>to</td><td>integrate</td></tr>
<tr><td>therefore</td><td>have</td><td>we</td><td colspan="3">every reason</td><td>the</td><td>environment</td><td colspan="2">in the</td><td>agricultural policy ,</td><td colspan="2">to integrate</td></tr>
<tr><td>that is why</td><td colspan="2">we have</td><td>all</td><td>reason</td><td>,</td><td>which</td><td colspan="2">environment in</td><td colspan="2">agricultural policy</td><td colspan="2">parliament</td></tr>
<tr><td colspan="2">have therefore</td><td>us</td><td>all the</td><td colspan="2">reason</td><td>of the</td><td colspan="2">environment into</td><td colspan="2">the agricultural policy</td><td colspan="2">successfully integrated</td></tr>
<tr><td colspan="2">hence</td><td>, we</td><td>every</td><td colspan="3">reason to make</td><td>environmental</td><td>on</td><td colspan="2">the cap</td><td colspan="2">be woven together</td></tr>
<tr><td colspan="3">we have therefore</td><td>everyone</td><td colspan="3">grounds for taking the</td><td>the environment</td><td colspan="2">to the</td><td>agricultural policy is</td><td>on</td><td>parliament</td></tr>
<tr><td>so</td><td colspan="2">, we</td><td>all of</td><td>cause</td><td colspan="2">which</td><td>environment ,</td><td>to</td><td colspan="2">the cap ,</td><td>for</td><td>incorporated</td></tr>
<tr><td colspan="3">hence our</td><td>any</td><td>why</td><td colspan="2">that</td><td>outside</td><td>at</td><td colspan="2">agricultural policy</td><td>too</td><td>woven together</td></tr>
<tr><td colspan="2">therefore ,</td><td>it</td><td>of all</td><td>reason for</td><td colspan="2">, the</td><td>completion</td><td>into</td><td colspan="2">that agricultural policy</td><td>be</td><td></td></tr>
</table>

And many, many more…even before reordering

Search in Phrase Models

<table>
<tr><th>Deshalb</th><th>haben</th><th>wir</th><th>allen</th><th>Grund</th><th>,</th><th>die</th><th>Umwelt</th><th>in</th><th>die</th><th>Agrarpolitik</th><th>zu</th><th>integrieren</th></tr>
<tr><td colspan="3">that is why we have</td><td colspan="2">every reason</td><td colspan="3">the environment</td><td>in</td><td>the</td><td>agricultural policy</td><td>to</td><td>integrate</td></tr>
<tr><td>therefore</td><td>have</td><td>we</td><td colspan="3">every reason</td><td>the</td><td>environment</td><td colspan="2">in the</td><td>agricultural policy ,</td><td colspan="2">to integrate</td></tr>
<tr><td>that is why</td><td colspan="2">we have</td><td>all</td><td>reason</td><td>,</td><td>which</td><td colspan="2">environment in</td><td colspan="2">agricultural policy</td><td colspan="2">parliament</td></tr>
<tr><td colspan="2">have therefore</td><td>us</td><td>all the</td><td colspan="2">reason</td><td>of the</td><td colspan="2">environment into</td><td colspan="2">the agricultural policy</td><td colspan="2">successfully integrated</td></tr>
<tr><td colspan="2">hence</td><td>, we</td><td>every</td><td colspan="3">reason to make</td><td>environmental</td><td>on</td><td colspan="2">the cap</td><td colspan="2">be woven together</td></tr>
<tr><td colspan="3">we have therefore</td><td>everyone</td><td colspan="3">grounds for taking the</td><td>the environment</td><td colspan="2">to the</td><td>agricultural policy is</td><td>on</td><td>parliament</td></tr>
<tr><td>so</td><td colspan="2">, we</td><td>all of</td><td>cause</td><td colspan="2">which</td><td>environment ,</td><td>to</td><td colspan="2">the cap ,</td><td>for</td><td>incorporated</td></tr>
<tr><td colspan="3">hence our</td><td>any</td><td>why</td><td colspan="2">that</td><td>outside</td><td>at</td><td colspan="2">agricultural policy</td><td>too</td><td>woven together</td></tr>
<tr><td colspan="2">therefore ,</td><td>it</td><td>of all</td><td>reason for</td><td colspan="2">, the</td><td>completion</td><td>into</td><td colspan="2">that agricultural policy</td><td>be</td><td></td></tr>
</table>

And many, many more…even before reordering

Search in Phrase Models

<table>
<tr><th>Deshalb</th><th>haben</th><th>wir</th><th>allen</th><th>Grund</th><th>,</th><th>die</th><th>Umwelt</th><th>in</th><th>die</th><th>Agrarpolitik</th><th>zu</th><th>integrieren</th></tr>
<tr><td colspan="3">that is why we have</td><td colspan="2">every reason</td><td colspan="3">the environment</td><td>in</td><td>the</td><td>agricultural policy</td><td>to</td><td>integrate</td></tr>
<tr><td>therefore</td><td>have</td><td>we</td><td colspan="3">every reason</td><td>the</td><td>environment</td><td colspan="2">in the</td><td>agricultural policy ,</td><td colspan="2">to integrate</td></tr>
<tr><td>that is why</td><td colspan="2">we have</td><td>all</td><td>reason</td><td>,</td><td>which</td><td colspan="2">environment in</td><td colspan="2">agricultural policy</td><td colspan="2">parliament</td></tr>
<tr><td colspan="2">have therefore</td><td>us</td><td>all the</td><td colspan="2">reason</td><td>of the</td><td colspan="2">environment into</td><td colspan="2">the agricultural policy</td><td colspan="2">successfully integrated</td></tr>
<tr><td colspan="2">hence</td><td>, we</td><td>every</td><td colspan="3">reason to make</td><td>environmental</td><td>on</td><td colspan="2">the cap</td><td colspan="2">be woven together</td></tr>
<tr><td colspan="3">we have therefore</td><td>everyone</td><td colspan="3">grounds for taking the</td><td>the environment</td><td colspan="2">to the</td><td>agricultural policy is</td><td>on</td><td>parliament</td></tr>
<tr><td>so</td><td colspan="2">, we</td><td>all of</td><td>cause</td><td colspan="2">which</td><td>environment ,</td><td>to</td><td colspan="2">the cap ,</td><td>for</td><td>incorporated</td></tr>
<tr><td colspan="3">hence our</td><td>any</td><td>why</td><td colspan="2">that</td><td>outside</td><td>at</td><td colspan="2">agricultural policy</td><td>too</td><td>woven together</td></tr>
<tr><td colspan="2">therefore ,</td><td>it</td><td>of all</td><td>reason for</td><td colspan="2">, the</td><td>completion</td><td>into</td><td colspan="2">that agricultural policy</td><td>be</td><td></td></tr>
</table>

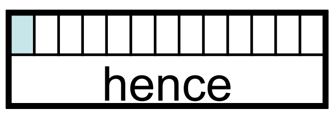
And many, many more…even before reordering

“Stack Decoding”

Deshalb haben wir allen Grund , die Umwelt in die Agrarpolitik zu integrieren

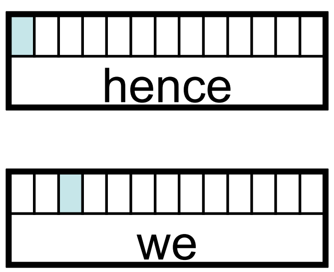
“Stack Decoding”

Deshalb haben wir allen Grund , die Umwelt in die Agrarpolitik zu integrieren



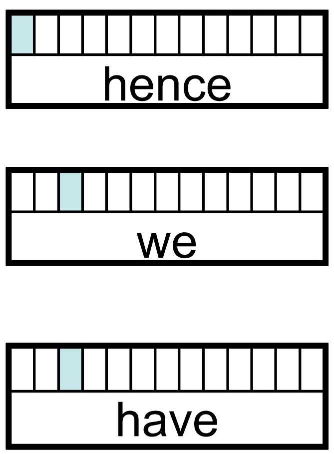
“Stack Decoding”

Deshalb haben wir allen Grund , die Umwelt in die Agrarpolitik zu integrieren



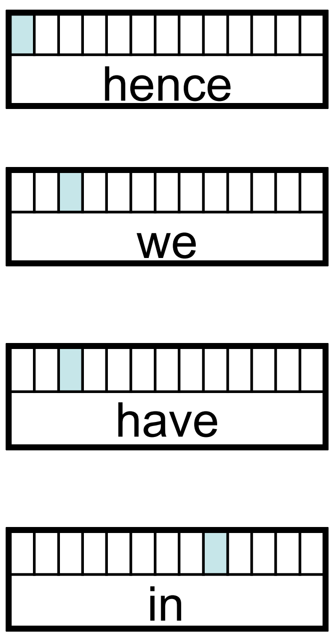
“Stack Decoding”

Deshalb haben wir allen Grund , die Umwelt in die Agrarpolitik zu integrieren



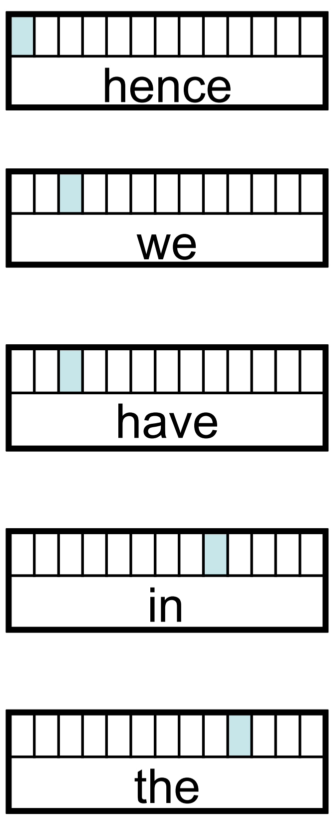
“Stack Decoding”

Deshalb haben wir allen Grund , die Umwelt in die Agrarpolitik zu integrieren



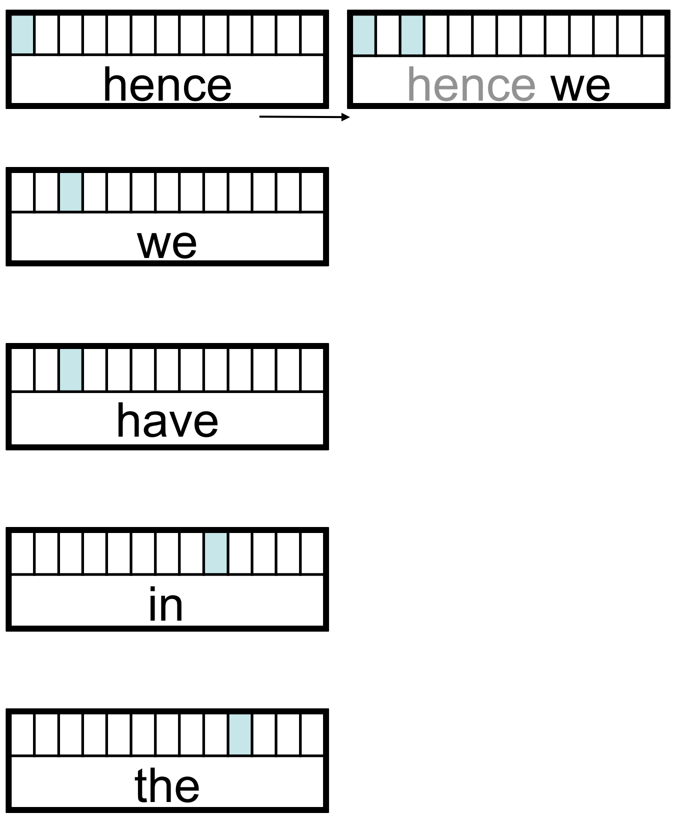
“Stack Decoding”

Deshalb haben wir allen Grund , die Umwelt in die Agrarpolitik zu integrieren



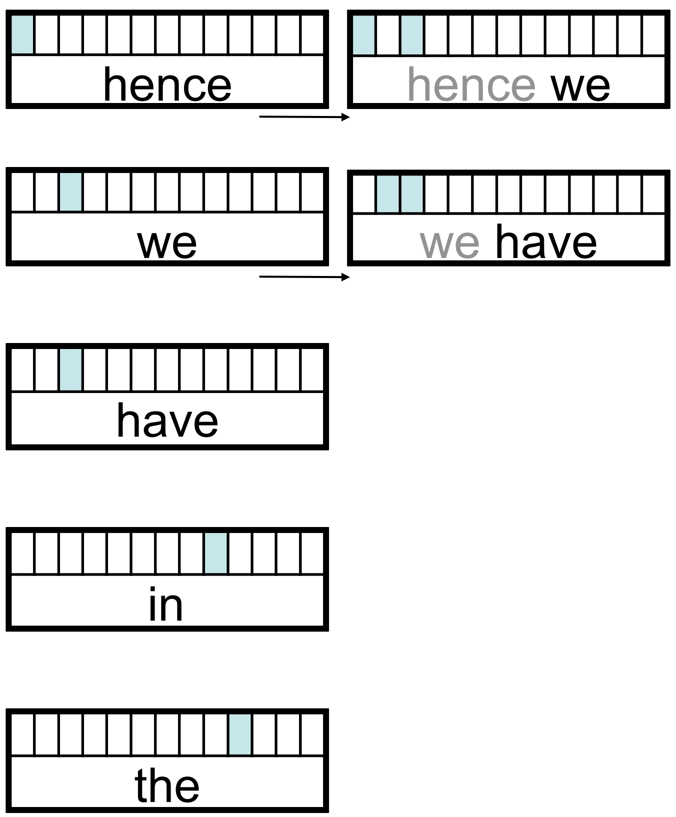
“Stack Decoding”

Deshalb haben wir allen Grund , die Umwelt in die Agrarpolitik zu integrieren



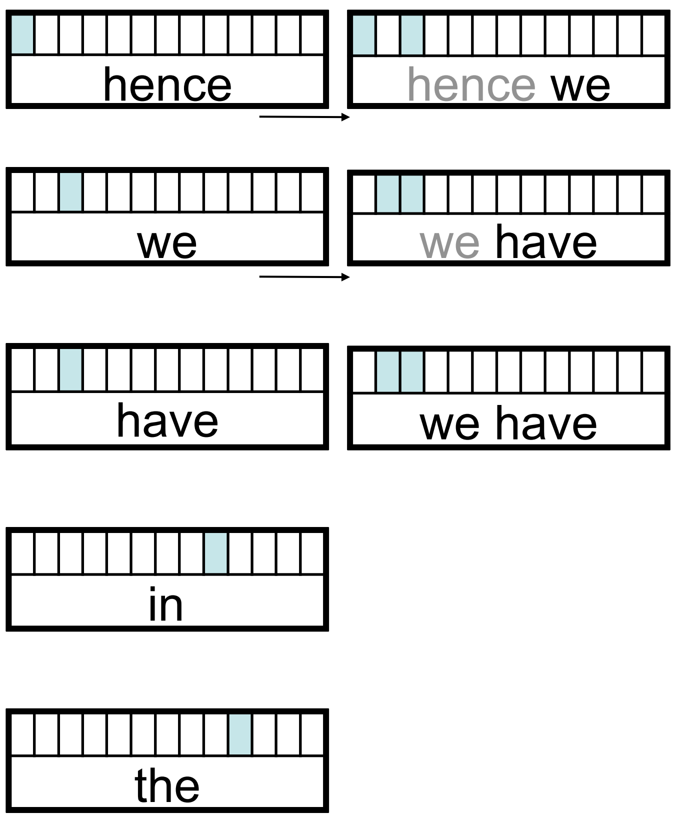
“Stack Decoding”

Deshalb haben wir allen Grund , die Umwelt in die Agrarpolitik zu integrieren



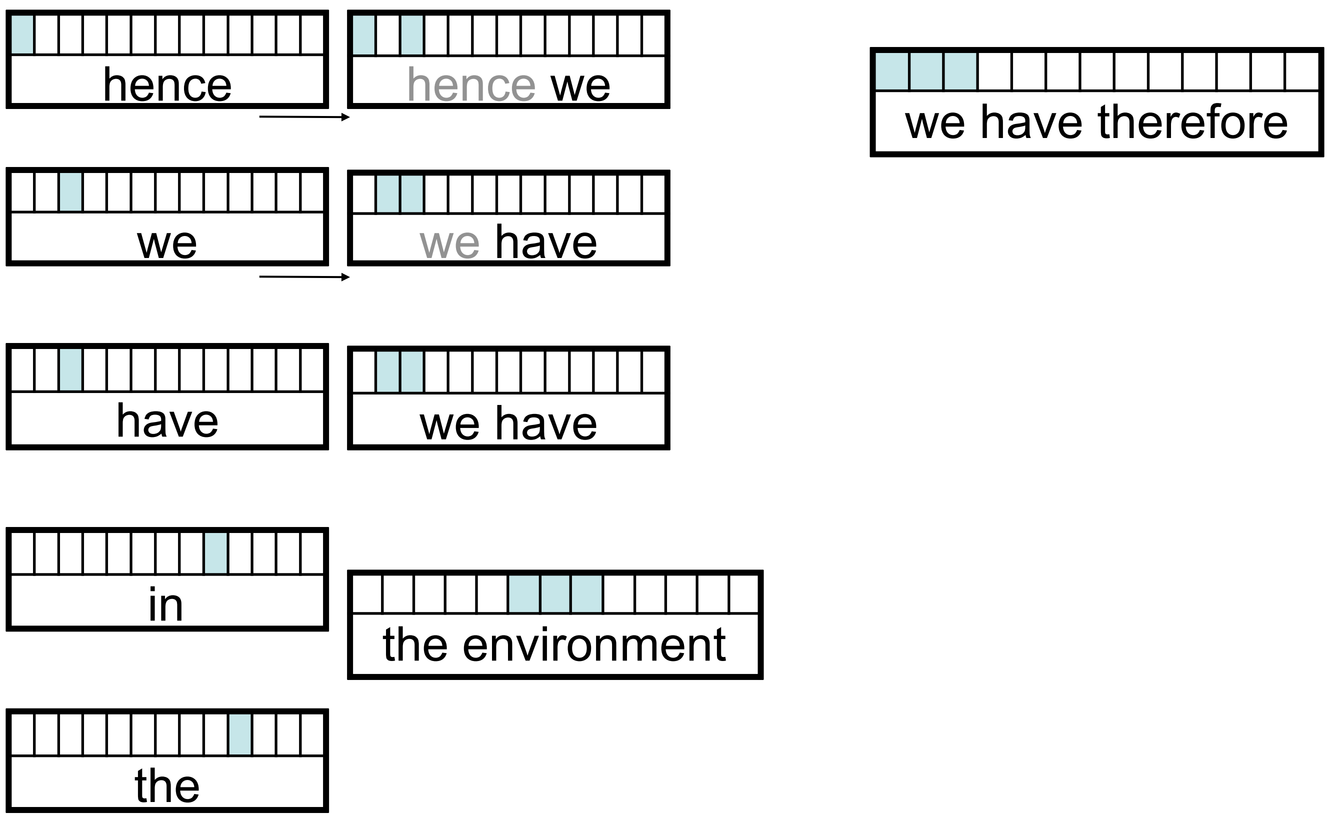
“Stack Decoding”

Deshalb haben wir allen Grund , die Umwelt in die Agrarpolitik zu integrieren



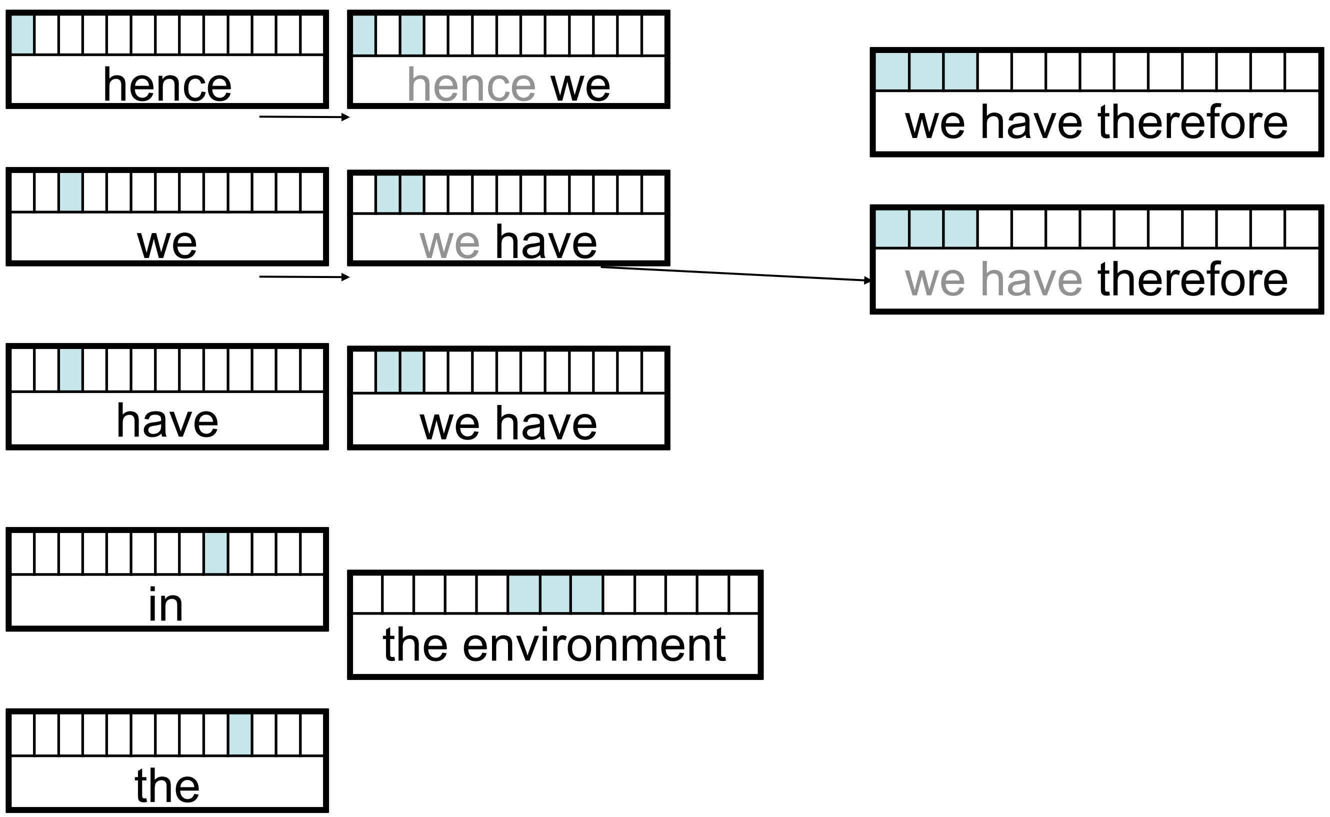
“Stack Decoding”

Deshalb haben wir allen Grund , die Umwelt in die Agrarpolitik zu integrieren



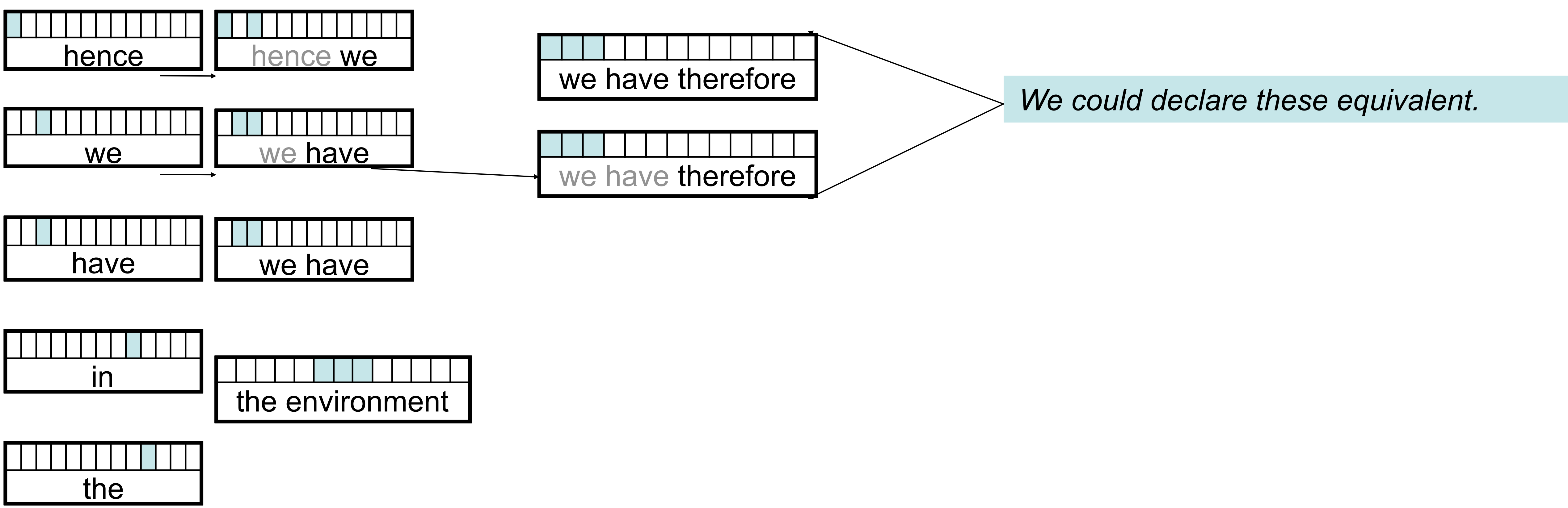
“Stack Decoding”

Deshalb haben wir allen Grund , die Umwelt in die Agrarpolitik zu integrieren



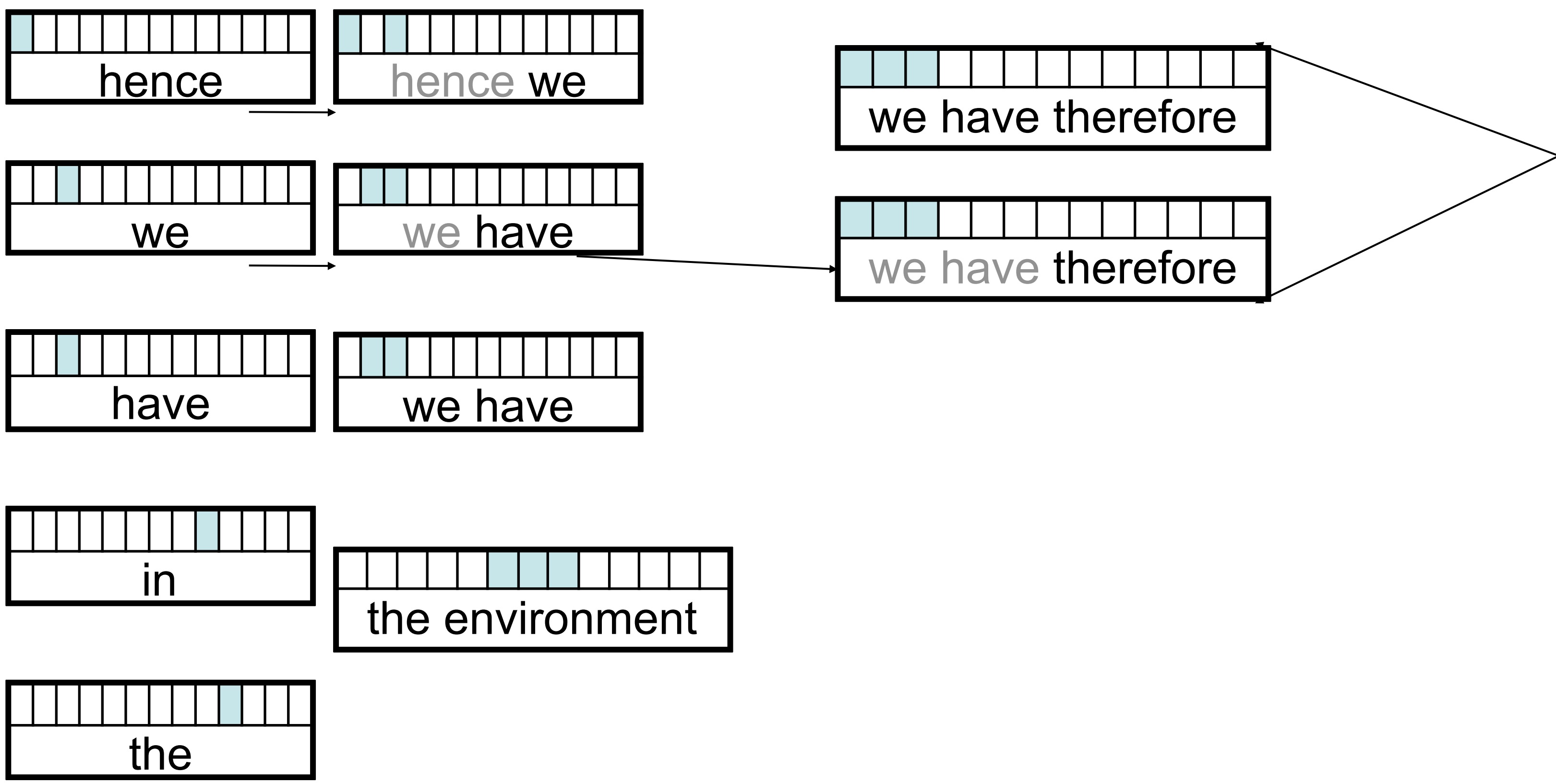
“Stack Decoding”

Deshalb haben wir allen Grund , die Umwelt in die Agrarpolitik zu integrieren



“Stack Decoding”

Deshalb haben wir allen Grund , die Umwelt in die Agrarpolitik zu integrieren



We could declare these equivalent.

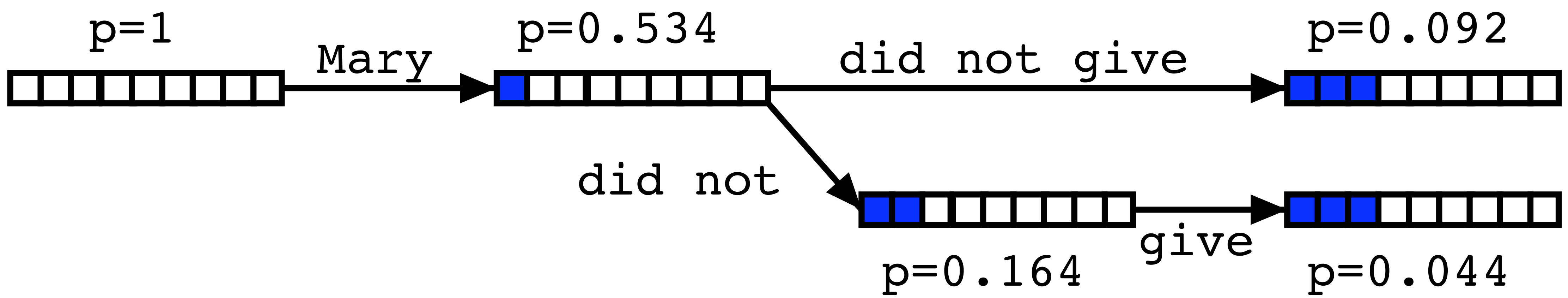
etc., u.s.w., until all source words are covered

Search in Phrase Models

- Many ways of segmenting source
- Many ways of translating each segment
- *Restrict* model class: phrases >, e.g., 7 words, no long-distance reordering
- *Recombine* equivalent hypotheses
- *Prune* away unpromising partial translations or we'll run out of space and/or run too long
  - –How to compare partial translations?
  - –Some start with easy stuff: “in”, “das”, ...
  - –Some with hard stuff: “Agrarpolitik”, “Entscheidungsproblem”, …

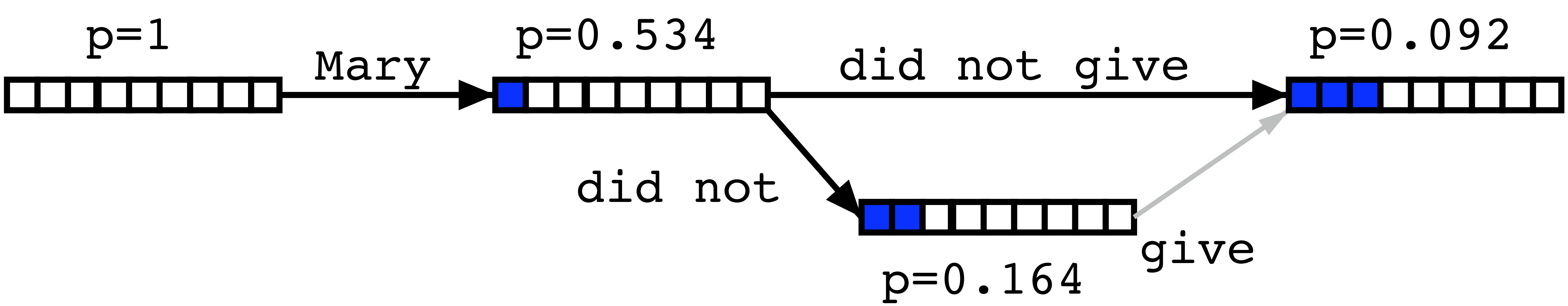
Hypothesis Recombination

- Different paths to the same partial translation



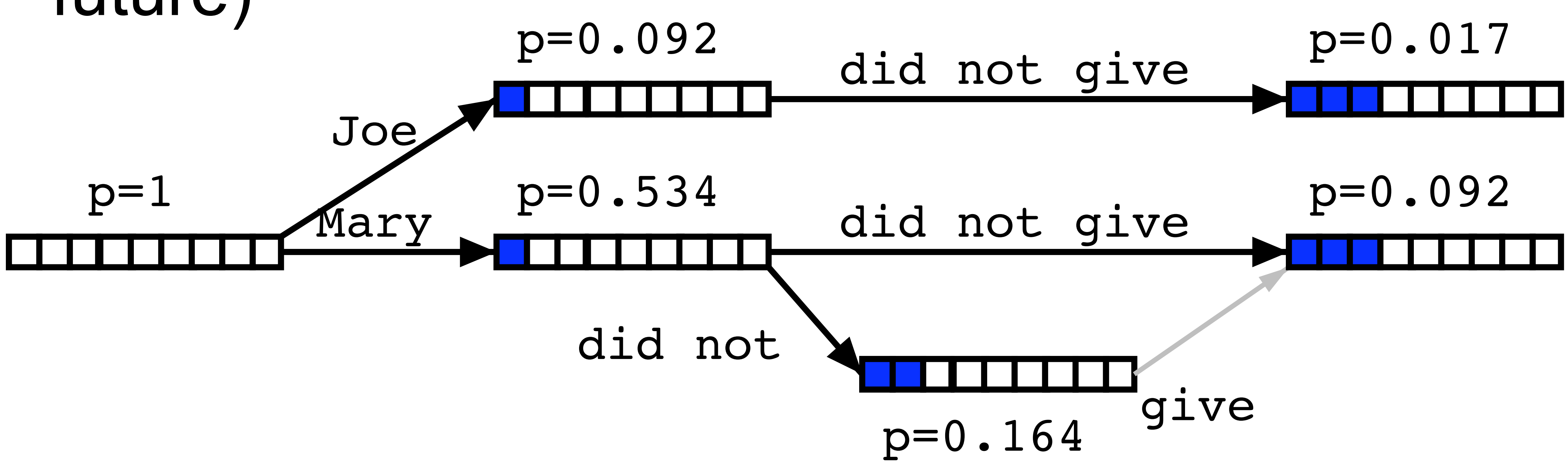
Hypothesis Recombination

- Different paths to the same partial translation
- Combine paths
  - Drop weaker path
  - Keep backpointer to weaker path (for lattice or n-best generation)



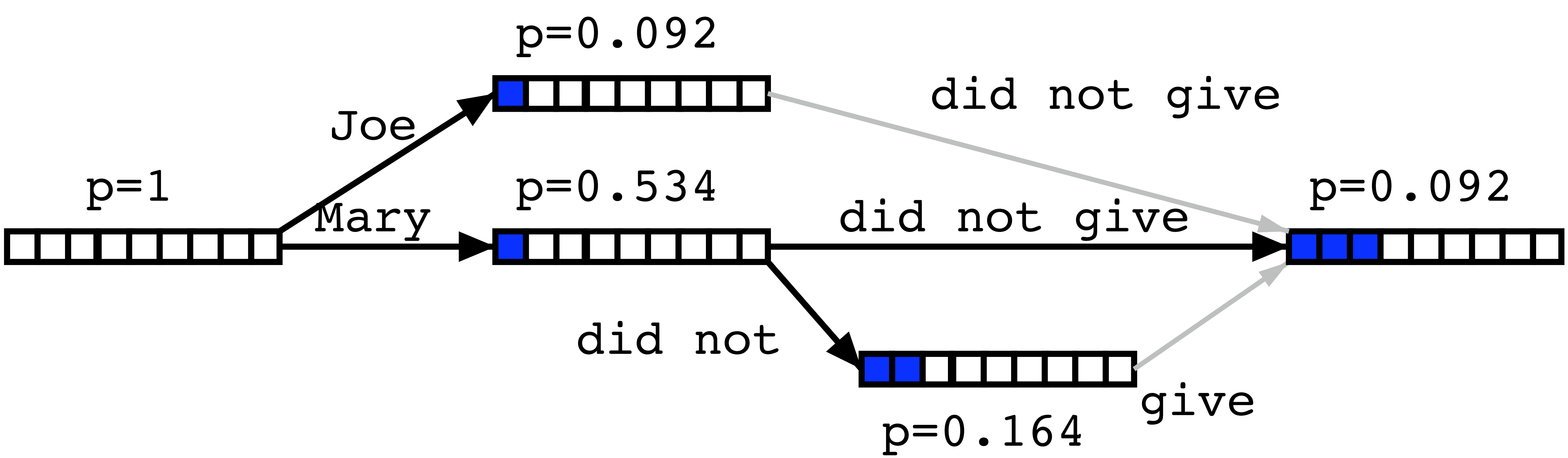
Hypothesis Recombination

- Recombined hypotheses do not have to match completely
- Weaker path can be dropped if
  - Last n target words match (for n+1-gram lang. model)
  - Source coverage vectors match (same best future)



Hypothesis Recombination

- Combining partially matching hypotheses



Pruning

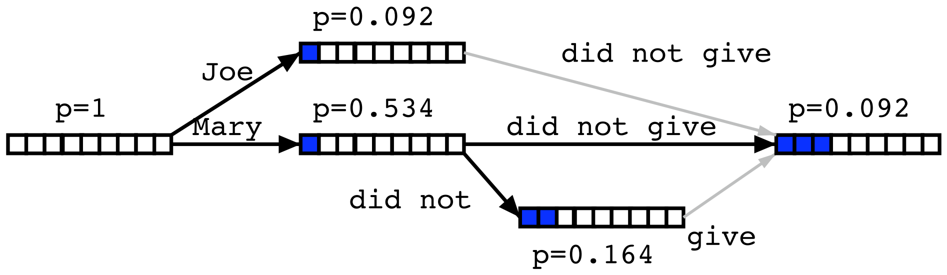
- Hypothesis recombination is *not su cient*

  Heuristically *discard* weak hypotheses early

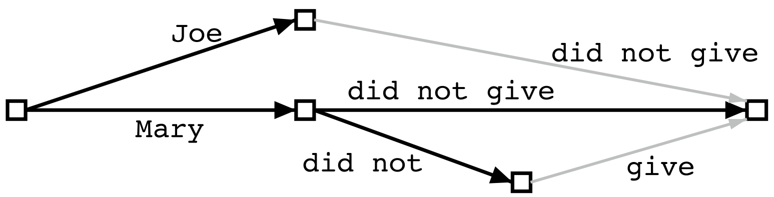
- Organize Hypothesis in **stacks**, e.g. by
  - *same* foreign words covered
  - *same number* of foreign words covered
  - *same number* of English words produced

- Compare hypotheses in stacks, discard bad ones
  - **histogram pruning**: keep top n hypotheses in each stack (e.g., $n = 100$)
  - **threshold pruning**: keep hypotheses that are at most times the cost of best hypothesis in stack (e.g., $= 0.001$)

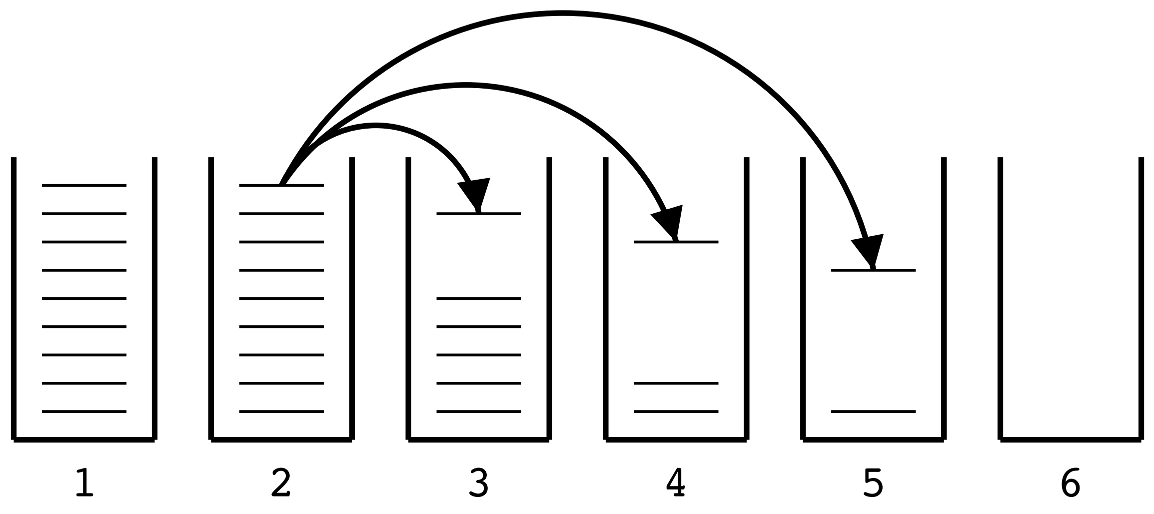
Word Lattice Generation



- **Search graph** can be easily converted into a **word lattice**
  - can be further mined for **n-best lists**
    enables **reranking** approaches
    enables **discriminative training**



Hypothesis Stacks



- Organization of hypothesis into stacks
  - here: based on *number of foreign words* translated
  - during translation all hypotheses from one stack are expanded
  - expanded Hypotheses are placed into stacks

Limits on Reordering

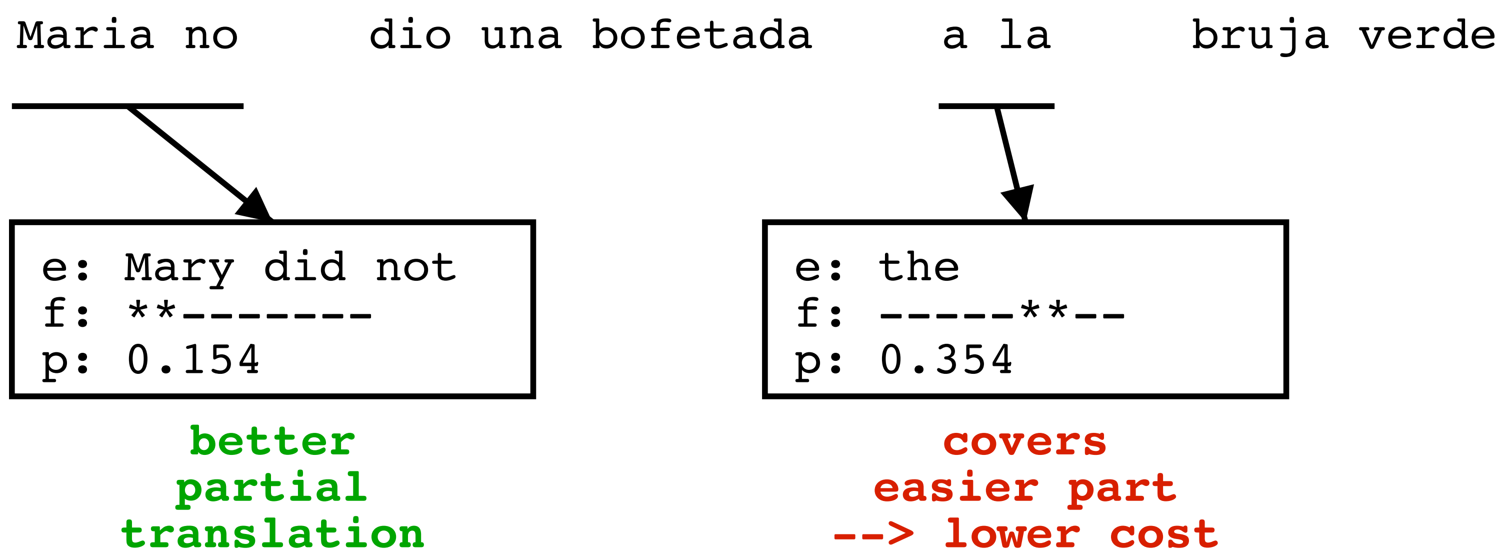
- Reordering may be **limited**
  - **Monotone** Translation: No reordering at all
  - Only phrase movements of at most n words

- Reordering limits *speed* up search (polynomial instead of exponential)

- Current reordering models are weak, so limits *improve* translation quality

Comparing Hypotheses

- Comparing hypotheses with *same number of foreign words* covered



- Hypothesis that covers *easy part* of sentence is preferred

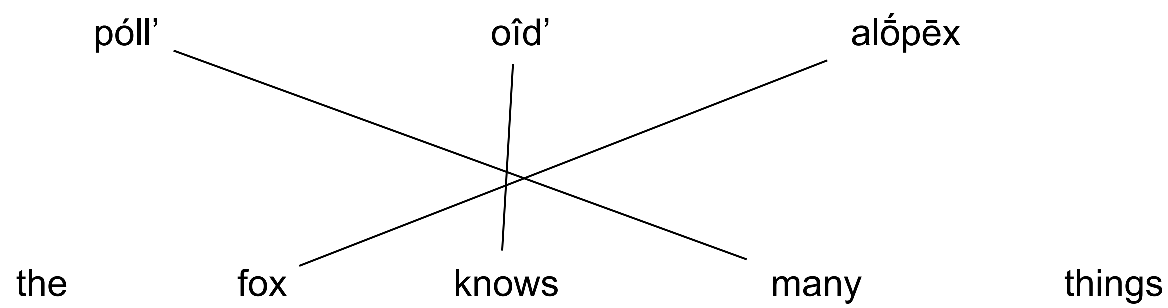
  Need to consider **future cost** of uncovered parts

  or: have one hypothesis stack per coverage vector

Synchronous Grammars

- Just like monolingual grammars except...
  - Each rule involves pairs (tuples) of nonterminals
  - Tuples of elementary trees for TAG, etc.
- First proposed for source-source translation in compilers
- Can be constituency, dependency, lexicalized, etc.
- Parsing speedups for monolingual grammar don't necessarily work
  - E.g., no split-head trick for lexicalized parsing
- Binarization less straightforward

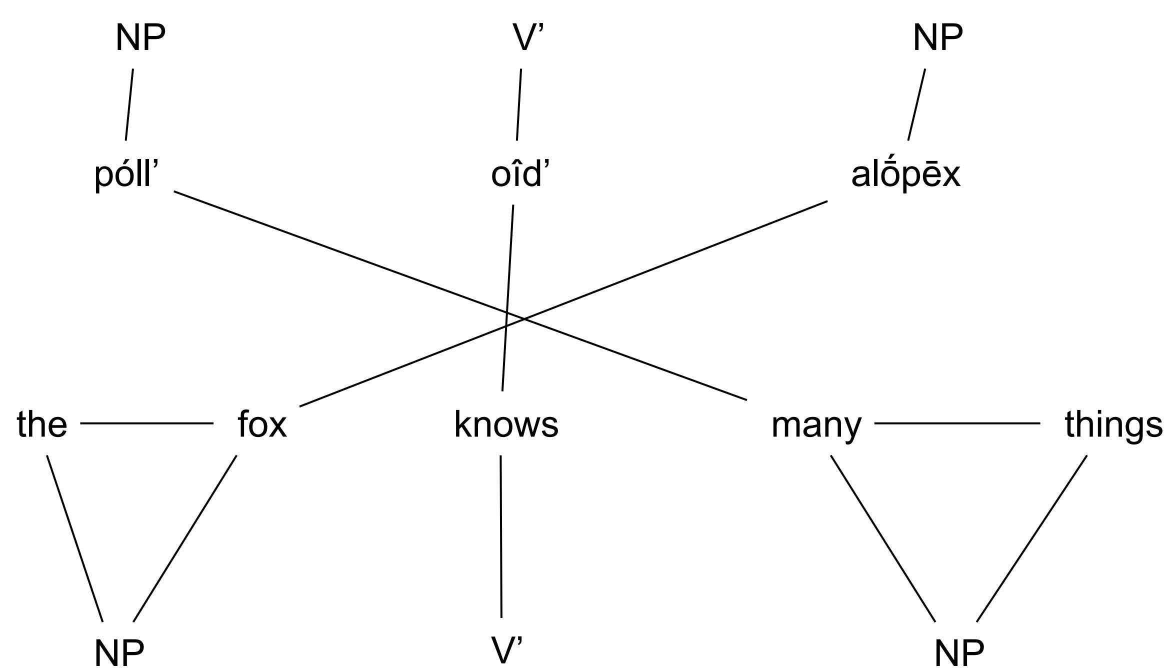
Bilingual Parsing



A variant of CKY chart parsing.

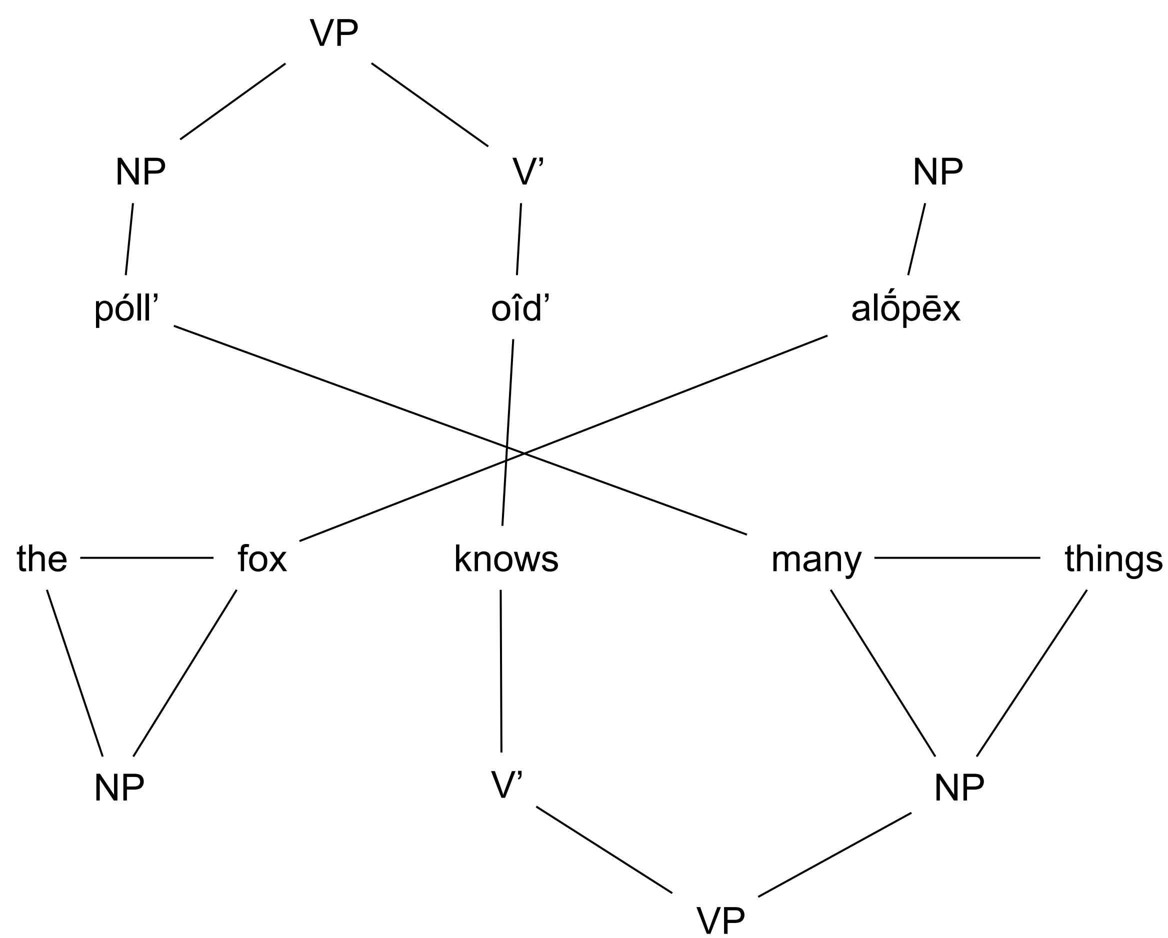
	póll’	oîd’	alṓpēx
the			
fox			NN/NN
knows		VB/VB	
many	JJ/JJ		
things			

Bilingual Parsing



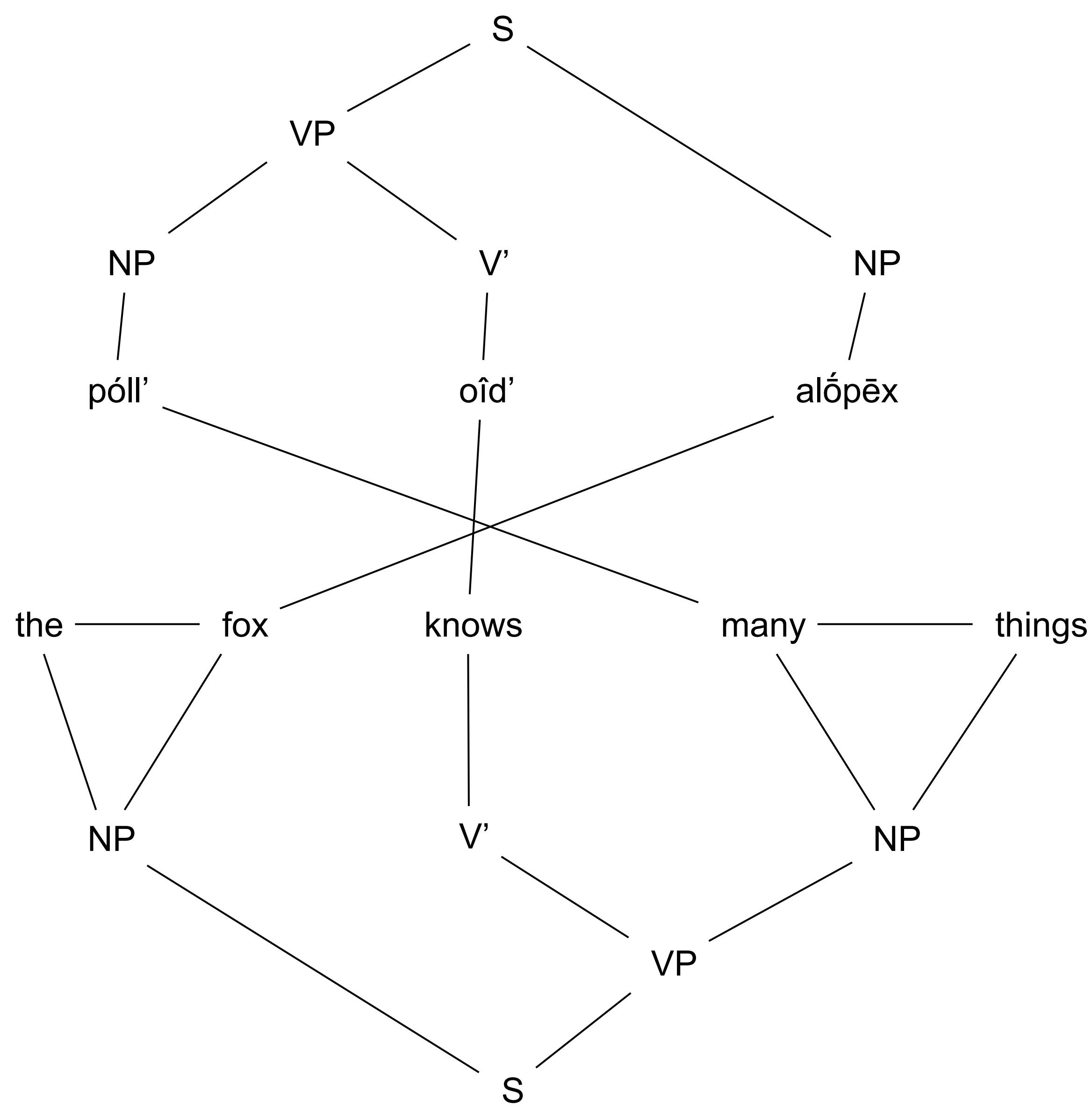
	póll’	oîd’	alṓpēx
the			NP/NP
fox			
knows		VP/VP	
many	NP/NP		
things			

Bilingual Parsing



	póll'	oîd'	alṓpēx
the			NP/NP
fox			
knows	VP/VP		
many			
things			

Bilingual Parsing



<table>
<tr><th></th><th>póll’</th><th>oîd’</th><th>alṓpēx</th></tr>
<tr><td>the</td><td rowspan="5" colspan="3">S/S</td></tr>
<tr><td>fox</td></tr>
<tr><td>knows</td></tr>
<tr><td>many</td></tr>
<tr><td>things</td></tr>
</table>

MT as Parsing

- If we only have the source, parse it while recording all compatible target language trees.
- Runtime is also multiplied by a *grammar constant*: one string could be a noun and a verb phrase
- Continuing problem of multiple hidden configurations (trees, instead of phrases) for one translation.

Parsing as Deduction

$$\forall A, B, C \in N, W \in V, 0 \leq i, j, k \leq m$$

$$constit(B, i, j) \wedge constit(C, j, k) \wedge A \rightarrow BC \Rightarrow constit(A, i, k)$$

$$word(W, i) \wedge A \rightarrow W \Rightarrow constit(A, i, i+1)$$

$$constit(A, i, k) = \bigvee_{B,C,j} constit(B, i, j) \wedge constit(C, j, k) \wedge A \rightarrow B\ C$$

$$constit(A, i, j) = \bigvee_{W} word(W, i, j) \wedge A \rightarrow W$$

Parsing as Deduction

$$constit(A,i,k) = \bigvee_{B,C,j} constit(B,i,j) \wedge constit(C,j,k) \wedge A \rightarrow B\ C$$

$$constit(A,i,j) = \bigvee_{W} word(W,i,j) \wedge A \rightarrow W$$

$$\begin{aligned} score(constit(A,i,k)) = \max_{B,C,j}\ & score(constit(B,i,j)) \\ & \cdot score(constit(C,j,k)) \\ & \cdot score(A \rightarrow B\ C) \end{aligned}$$

$$score(constit(A,i,j)) = \max_{W} score(word(W,i,j)) \cdot score(A \rightarrow W)$$

And how about the inside algorithm?

Bilingual Parsing: ITG

$$s(X,i,k,u,w) = \bigvee_{j,v,Y,Z} s(Y,i,j,u,v) \wedge s(Z,j,k,v,w) \wedge [X \rightarrow Y\ Z]$$

$$s(X,i,k,u,w) = \bigvee_{j,v,Y,Z} s(Y,i,j,v,w) \wedge s(Z,j,k,u,v) \wedge \langle X \rightarrow Y\ Z \rangle$$

$$s(X,i,j,u,v) = w(S,i,j) \wedge w(T,u,v) \wedge X \rightarrow S/T$$

$$s(X,i,j,u,u) = w(S,i,j) \wedge X \rightarrow S/\epsilon$$

$$s(X,i,i,u,v) = w(T,u,v) \wedge X \rightarrow \epsilon/T$$

Similar extensions for finding the best alignment or the expectations of particular alignments

What Makes Search Hard?

- What we really want: the best (highest-scoring) translation
- What we get: the best translation/phrase segmentation/alignment
  - Even summing over all ways of segmenting *one* translation is hard.
- Most common approaches:
  - Ignore problem
  - Sum over top *j* translation/segmentation/alignment triples to get top *k*<<j translations

Redundancy in *n*-best Lists

Source: Da ich wenig Zeit habe , gehe ich sofort in medias res .

as i have little time , i am immediately in medias res . | 0-1,0-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am immediately in medias res . | 0-0,0-0 1-1,1-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am in medias res immediately . | 0-1,0-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,9-9 9-9,10-10 10-10,11-11 11-11,8-8 12-12,12-12
as i have little time , i am in medias res immediately . | 0-0,0-0 1-1,1-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,9-9 9-9,10-10 10-10,11-11 11-11,8-8 12-12,12-12
as i have little time , i am immediately in medias res . | 0-1,0-1 2-2,4-4 3-3,2-2 4-4,3-3 5-5,5-5 6-7,6-7 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am immediately in medias res . | 0-0,0-0 1-1,1-1 2-2,4-4 3-3,2-2 4-4,3-3 5-5,5-5 6-7,6-7 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am in medias res immediately . | 0-1,0-1 2-2,4-4 3-3,2-2 4-4,3-3 5-5,5-5 6-7,6-7 8-8,9-9 9-9,10-10 10-10,11-11 11-11,8-8 12-12,12-12
as i have little time , i am in medias res immediately . | 0-0,0-0 1-1,1-1 2-2,4-4 3-3,2-2 4-4,3-3 5-5,5-5 6-7,6-7 8-8,9-9 9-9,10-10 10-10,11-11 11-11,8-8 12-12,12-12
as i have little time , i am immediately in medias res . | 0-1,0-1 2-2,4-4 3-4,2-3 5-5,5-5 6-6,7-7 7-7,6-6 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am immediately in medias res . | 0-0,0-0 1-1,1-1 2-2,4-4 3-4,2-3 5-5,5-5 6-6,7-7 7-7,6-6 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i would immediately in medias res . | 0-1,0-1 2-2,4-4 3-4,2-3 5-5,5-5 6-6,7-7 7-7,6-6 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
because i have little time , i am immediately in medias res . | 0-0,0-0 1-1,1-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am immediately in medias res . | 0-1,0-1 2-2,4-4 3-3,2-2 4-4,3-3 5-5,5-5 6-6,7-7 7-7,6-6 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am immediately in medias res . | 0-0,0-0 1-1,1-1 2-2,4-4 3-3,2-2 4-4,3-3 5-5,5-5 6-6,7-7 7-7,6-6 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am in res medias immediately . | 0-1,0-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,9-9 9-9,11-11 10-10,10-10 11-11,8-8 12-12,12-12
because i have little time , i am immediately in medias res . | 0-1,0-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am in res medias immediately . | 0-0,0-0 1-1,1-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,9-9 9-9,11-11 10-10,10-10 11-11,8-8 12-12,12-12

Training

Which features of data predict good translations?

Training: Generative/Discriminative

- Generative
  - –Maximum likelihood training: max p(data)
  - –“Count and normalize”
  - –Maximum likelihood with hidden structure
    - Expectation Maximization (EM)
- Discriminative training
  - –Maximum conditional likelihood
  - –Minimum error/risk training
  - –Other criteria: perceptron and max. margin

“Count and Normalize”

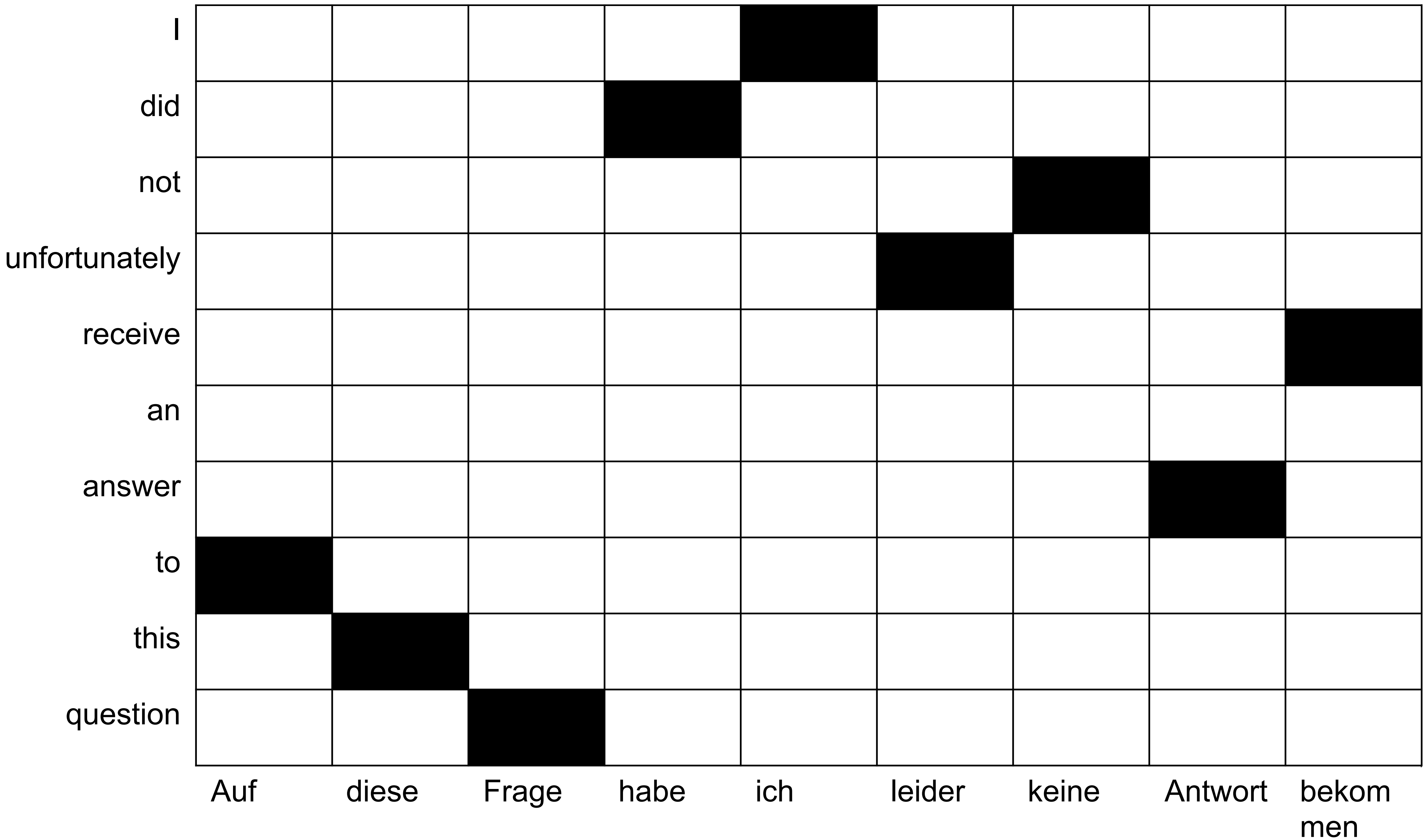
- Language modeling example: assume the probability of a word depends only on the previous 2 words.

$$p(\text{disease} \mid \text{into the}) = \frac{p(\text{into the disease})}{p(\text{into the})}$$

- p(disease|into the) = 3/20 = 0.15
- “Smoothing” reflects a prior belief that p(breech|into the) > 0 despite these 20 examples.

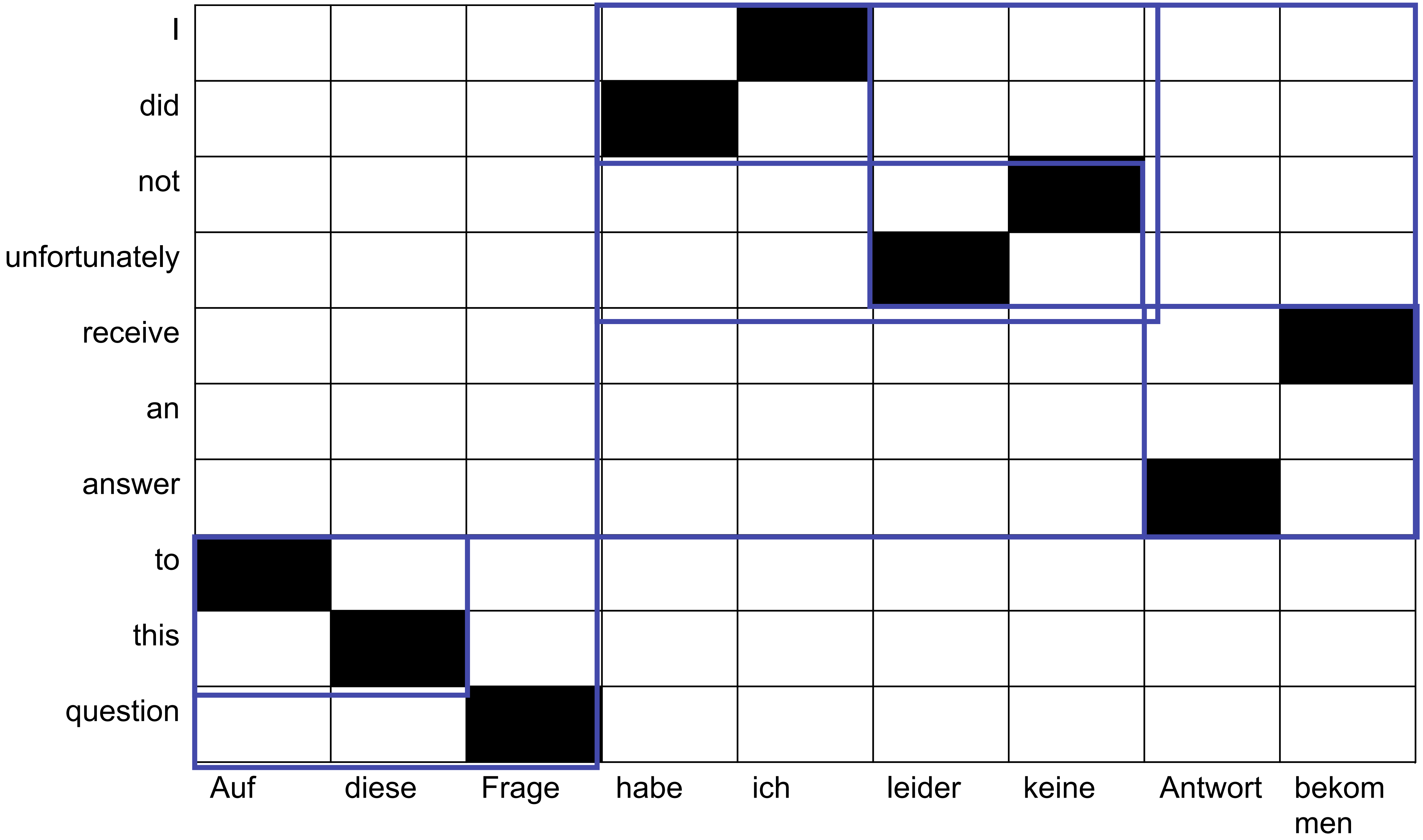
... into the programme ...
... into the **disease** ...
... into the **disease** ...
... into the correct ...
... into the next ...
... into the national ...
... into the integration ...
... into the Union ...
... into the Union ...
... into the Union ...
... into the sort ...
... into the internal ...
... into the general ...
... into the budget ...
... into the **disease** ...
... into the legal …
... into the various ...
... into the nuclear ...
... into the bargain ...
... into the situation ...

Phrase Models



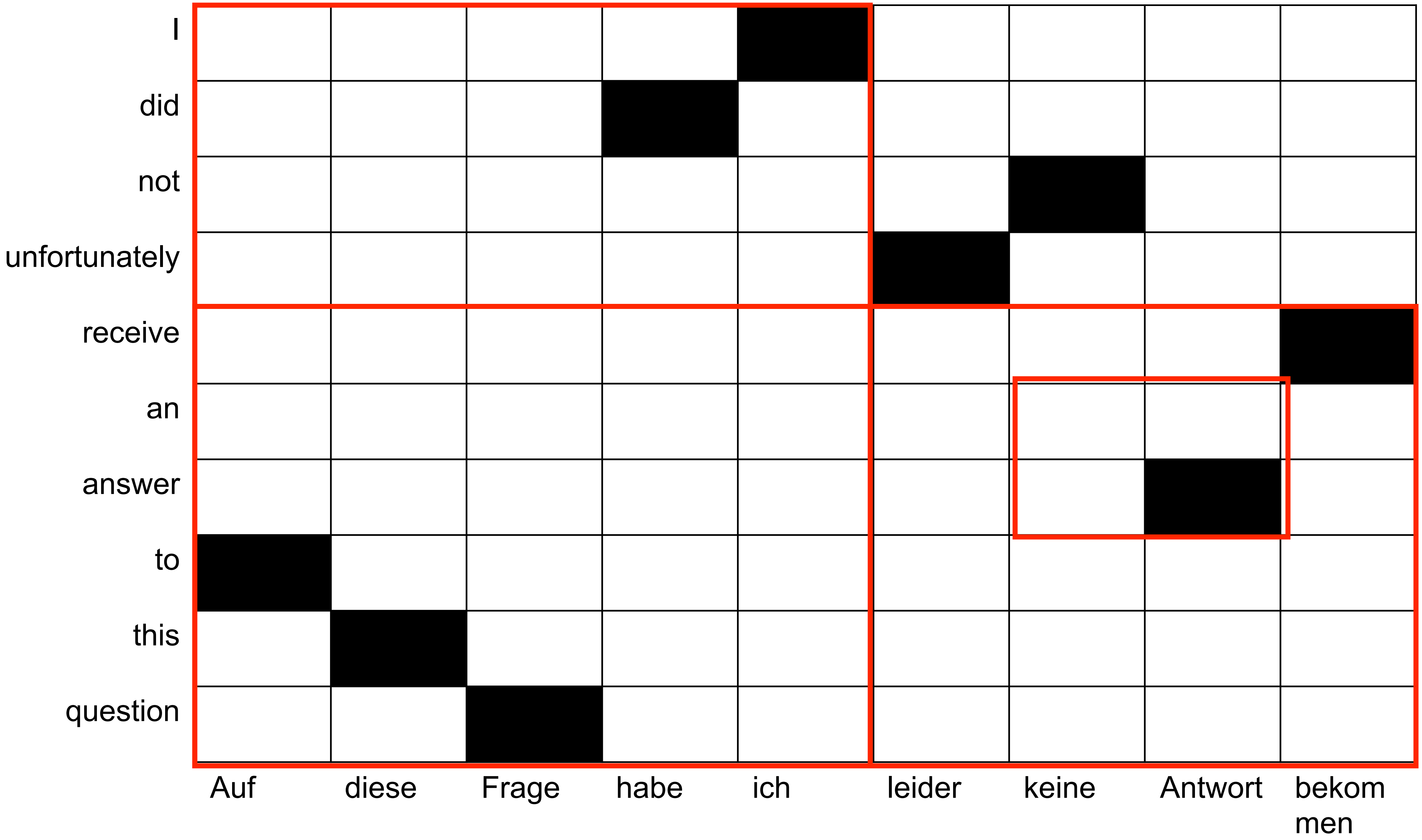
Assume word alignments are given.

Phrase Models



Some good phrase pairs.

Phrase Models



Some bad phrase pairs.

“Count and Normalize”

- Usual approach: treat relative frequencies of source phrase *s* and target phrase *t* as probabilities

$$p(s \mid t) \equiv \frac{count(s,t)}{count(t)} \qquad p(t \mid s) \equiv \frac{count(s,t)}{count(s)}$$

- This leads to overcounting when not all segmentations are legal due to unaligned words.

Hidden Structure

- But really, we don't observe word alignments.
- How are word alignment model parameters estimated?
- Find (all) structures consistent with observed data.
  - Some links are incompatible with others.
  - We need to score complete sets of links.

Hidden Structure and EM

- Expectation Maximization
  - Initialize model parameters (randomly, by some simpler model, or otherwise)
  - Calculate probabilities of hidden structures
  - Adjust parameters to maximize likelihood of observed data given hidden data
  - Iterate
- Summing over *all* hidden structures can be expensive
  - Sum over 1-best, *k*-best, other sampling methods

Discriminative Training

- Given a source sentence, give “good” translations a higher score than “bad” translations.
- We care about good translations, not a high probability of the training data.
- Spend less “energy” modeling bad translations.
- Disadvantages
  - We need to run the translation system at each training step.
  - System is tuned for one task (e.g. translation) and can’t be directly used for others (e.g. alignment)

“Good” Compared to What?

- Compare current translation to
- Idea #1: a human translation. OK, but
  - Good translations can be very dissimilar
  - We’d need to find hidden features (e.g. alignments)
- Idea #2: other top *n* translations (the “n-best list”). Better in practice, but
  - Many entries in n-best list are the same apart from hidden links
- Compare with a **loss function** *L*
  - 0/1: wrong or right; equal to reference or not
  - Task-specific metrics (word error rate, BLEU, …)

MT Evaluation

* **Intrinsic**

  Human evaluation

  Automatic (machine) evaluation

* **Extrinsic**

  How useful is MT system output for…

  Deciding whether a foreign language blog is about politics?
  Cross-language information retrieval?
  Flagging news stories about terrorist attacks?
  …

Human Evaluation

Je suis fatigué.

	Adequacy	Fluency
Tired is I.	5	2
Cookies taste good!	1	5
I am exhausted.	5	5

Human Evaluation

PRO

High quality

CON

Expensive!

Person (preferably bilingual) must make a time-consuming judgment per system hypothesis.

Expense prohibits frequent evaluation of incremental system modifications.

Automatic Evaluation

PRO

Cheap. Given available reference translations, free thereafter.

CON

We can only measure some proxy for translation quality.
(Such as N-Gram overlap or edit distance).

Output of Chinese-English system

In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong's export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province's total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong's export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province's total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion us dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Partially excellent translations

In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong's export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province's total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong's export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province's total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion US dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Mangled grammar

In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong's export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province's total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong's export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province's total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion us dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Evaluation of Machine Translation Systems

Bleu (Papineni, Roukos, Ward and Zhu, 2002):

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

Unigram Precision

- **Unigram Precision** of a candidate translation:

$$\frac{C}{N}$$

where N is number of words in the candidate, C is the number of words in the candidate which are in at least one reference translation.

e.g.,

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

$$Precision = \frac{17}{18}$$

(only *obeys* is missing from all reference translations)

Modified Unigram Precision

- Problem with unigram precision:

  Candidate: the the the the the the the

  Reference 1: the cat sat on the mat

  Reference 2: there is a cat on the mat

  precision = 7/7 = 1???

- **Modified unigram precision:** "Clipping"
  - Each word has a "cap". e.g., *cap(the) = 2*
  - A candidate word w can only be correct a maximum of $cap(w)$ times. e.g., in candidate above, $cap(the) = 2$, and *the* is correct twice in the candidate ⇒

$$Precision = \frac{2}{7}$$

Modified N-gram Precision

- Can generalize modified unigram precision to other n-grams.
- For example, for candidates 1 and 2 above:

$$Precision_1(bigram) = \frac{10}{17}$$

$$Precision_2(bigram) = \frac{1}{13}$$

Precision Alone Isn't Enough

Candidate 1: of the

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

$$Precision(unigram) = 1$$

$$Precision(bigram) = 1$$

But Recall isn't Useful in this Case

- Standard measure used in addition to precision is **recall**:

$$Recall = \frac{C}{N}$$

  where C is number of n-grams in candidate that are correct, N is number of words in the references.

  Candidate 1: I always invariably perpetually do.

  Candidate 2: I always do

  Reference 1: I always do

  Reference 1: I invariably do

  Reference 1: I perpetually do

Sentence Brevity Penalty

- Step 1: for each candidate, compute closest matching reference (in terms of length)
  e.g., our candidate is length 12, references are length $12, 15, 17$. Best match is of length 12.

- Step 2: Say l_i is the length of the i'th candidate, r_i is length of best match for the i'th candidate, then compute

$$brevity = \frac{\sum_i r_i}{\sum_i l_i}$$

  (I think! from the Papineni paper, although $brevity = \frac{\sum_i r_i}{\sum_i min(l_i, r_i)}$ might make more sense?)

- Step 3: compute brevity penalty

$$BP = \begin{cases} 1 & \text{If } brevity < 1 \\ e^{1-brevity} & \text{If } brevity \geq 1 \end{cases}$$

  e.g., if $r_i = 1.1 \times l_i$ for all i (candidates are always 10% too short) then $BP = e^{-0.1} = 0.905$

The Final Score

- Corpus precision for any n-gram is

$$p_n = \frac{\sum_{C \in \{Candidate\}} \sum_{ngram \in C} \quad Count_{clip}(ngram)}{\sum_{C \in \{Candidate\}} \sum_{ngram \in C} \quad Count(ngram)}$$

  i.e. number of correct ngrams in the candidates (after "clipping") divided by total number of ngrams in the candidates

- Final score is then

$$Bleu = BP \times (p_1 p_2 p_3 p_4)^{1/4}$$

  i.e., BP multiplied by the geometric mean of the unigram, bigram, trigram, and four-gram precisions

Automatic Evaluation: Bleu Score

hypothesis 1	I am exhausted
hypothesis 2	Tired is I
reference 1	I am tired
reference 2	I am ready to sleep now

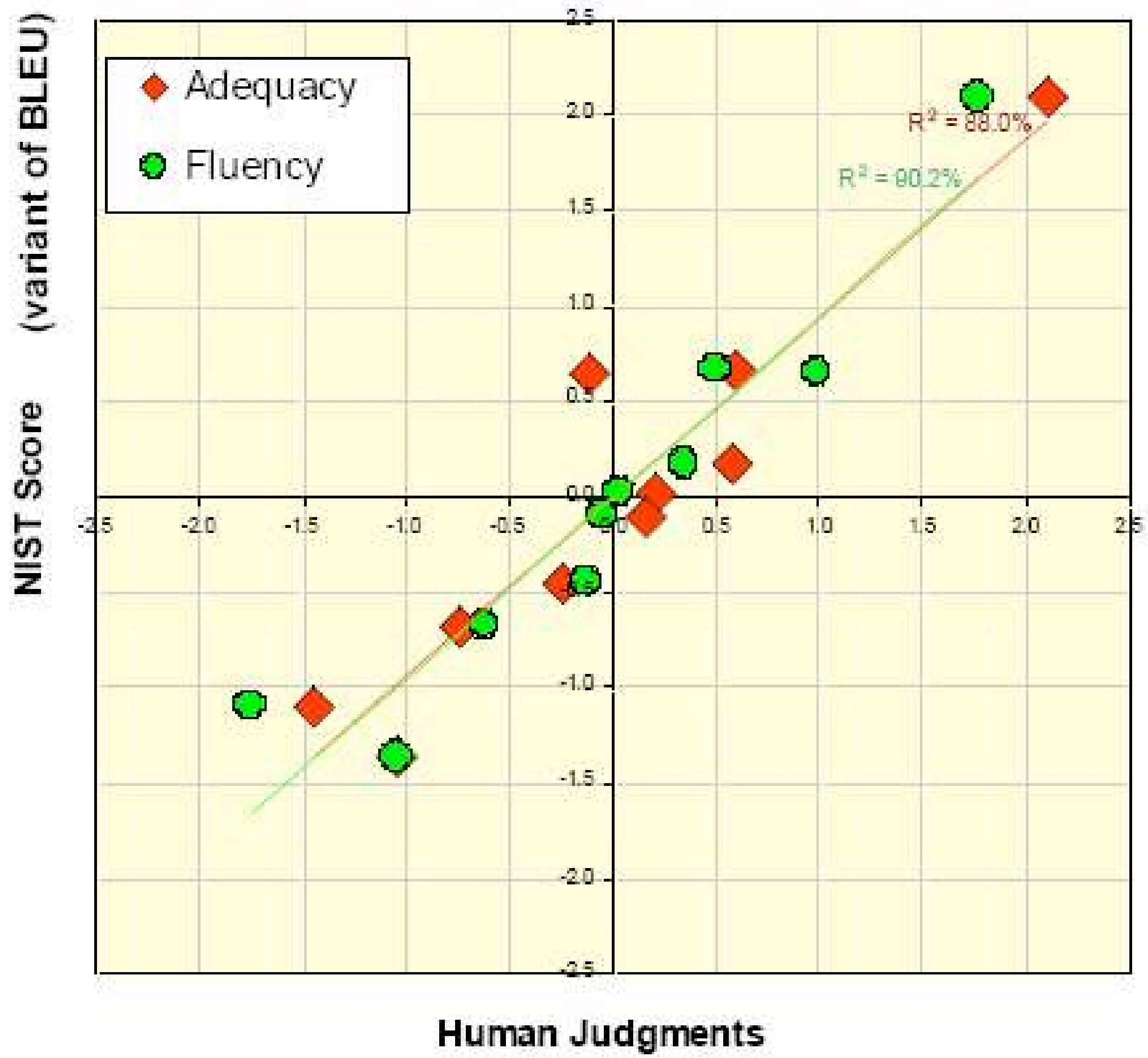
Automatic Evaluation: Bleu Score

		1-gram	2-gram	3-gram
hypothesis 1	I am exhausted	3/3	1/2	0/1
hypothesis 2	Tired is I	1/3	0/2	0/1
hypothesis 3	I I I	1/3	0/2	0/1

reference 1 I am tired

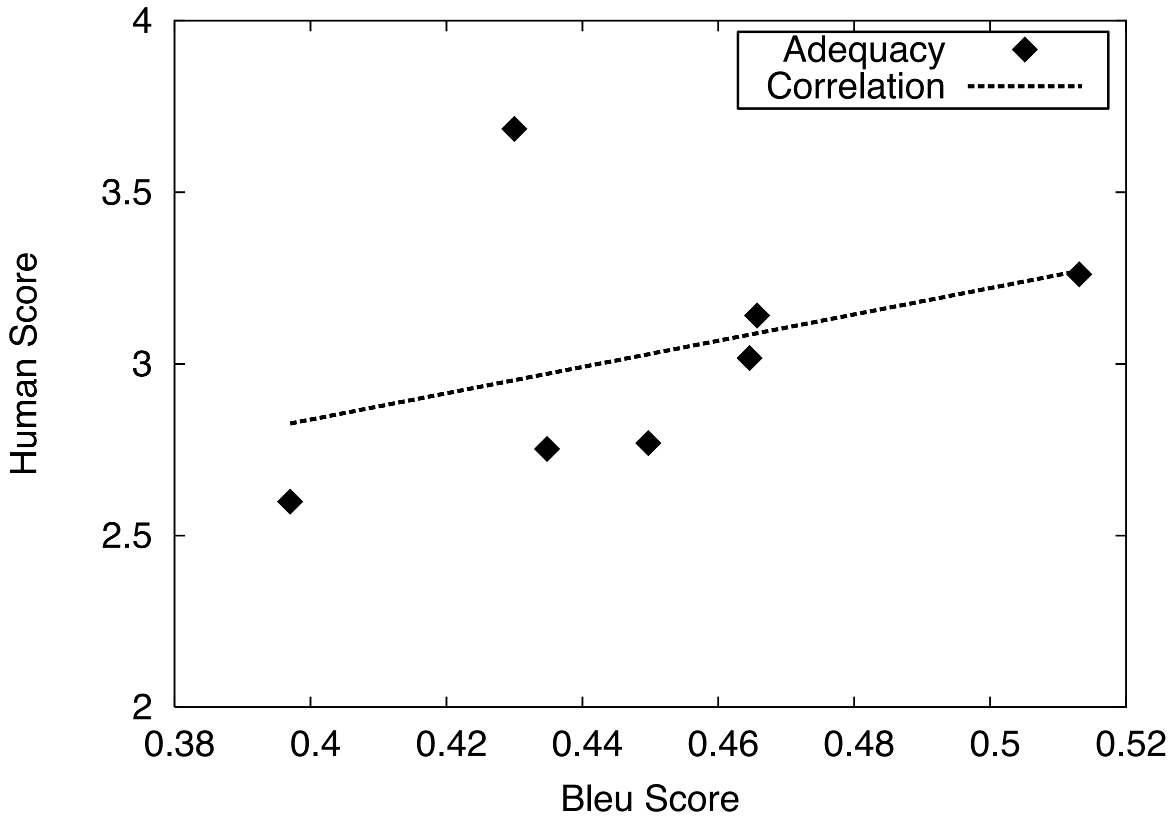
reference 2 I am ready to sleep now and so exhausted

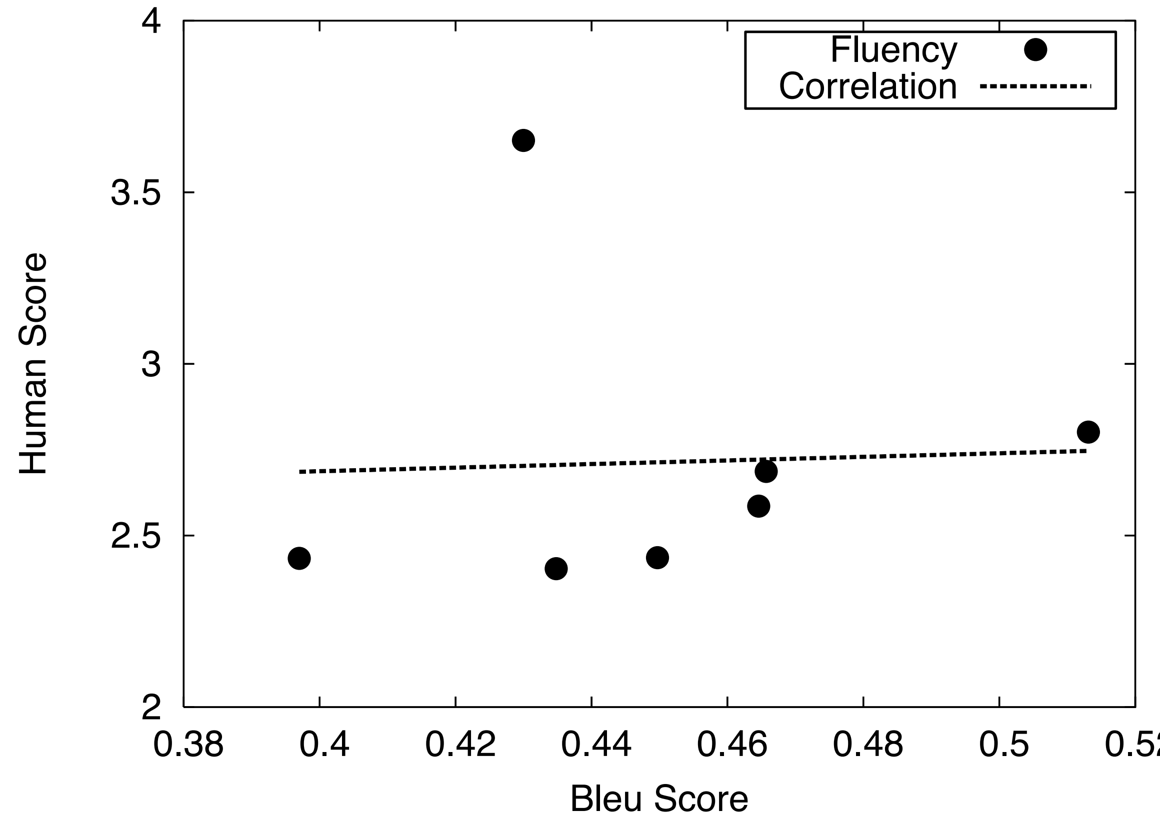
How Good are Automatic Metrics?



slide from G. Doddington (NIST)

Correlation? [Callison-Burch et al., 2006]



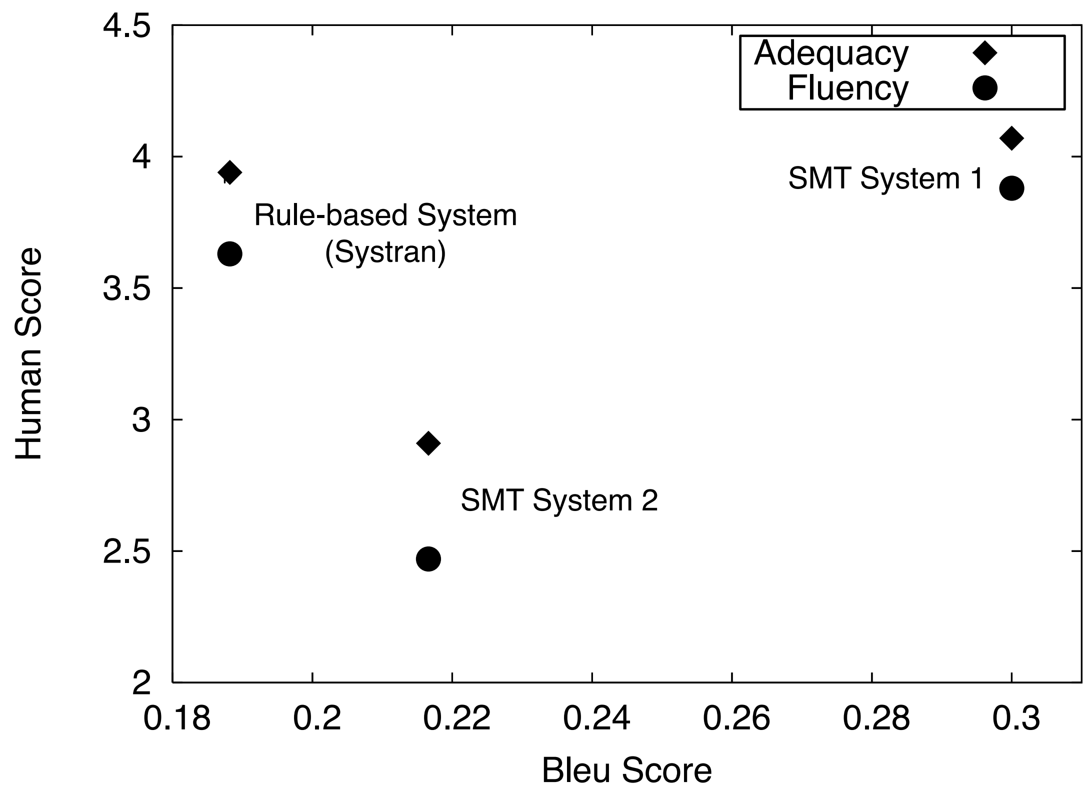


[from Callison-Burch et al., 2006, EACL]

- DARPA/NIST MT Eval 2005
  - Mostly statistical systems (all but one in graphs)
  - One submission **manual post-edit** of statistical system's output

$\rightarrow$ Good adequacy/fluency scores *not reflected* by BLEU

Correlation? [Callison-Burch et al., 2006]



[from Callison-Burch et al., 2006, EACL]

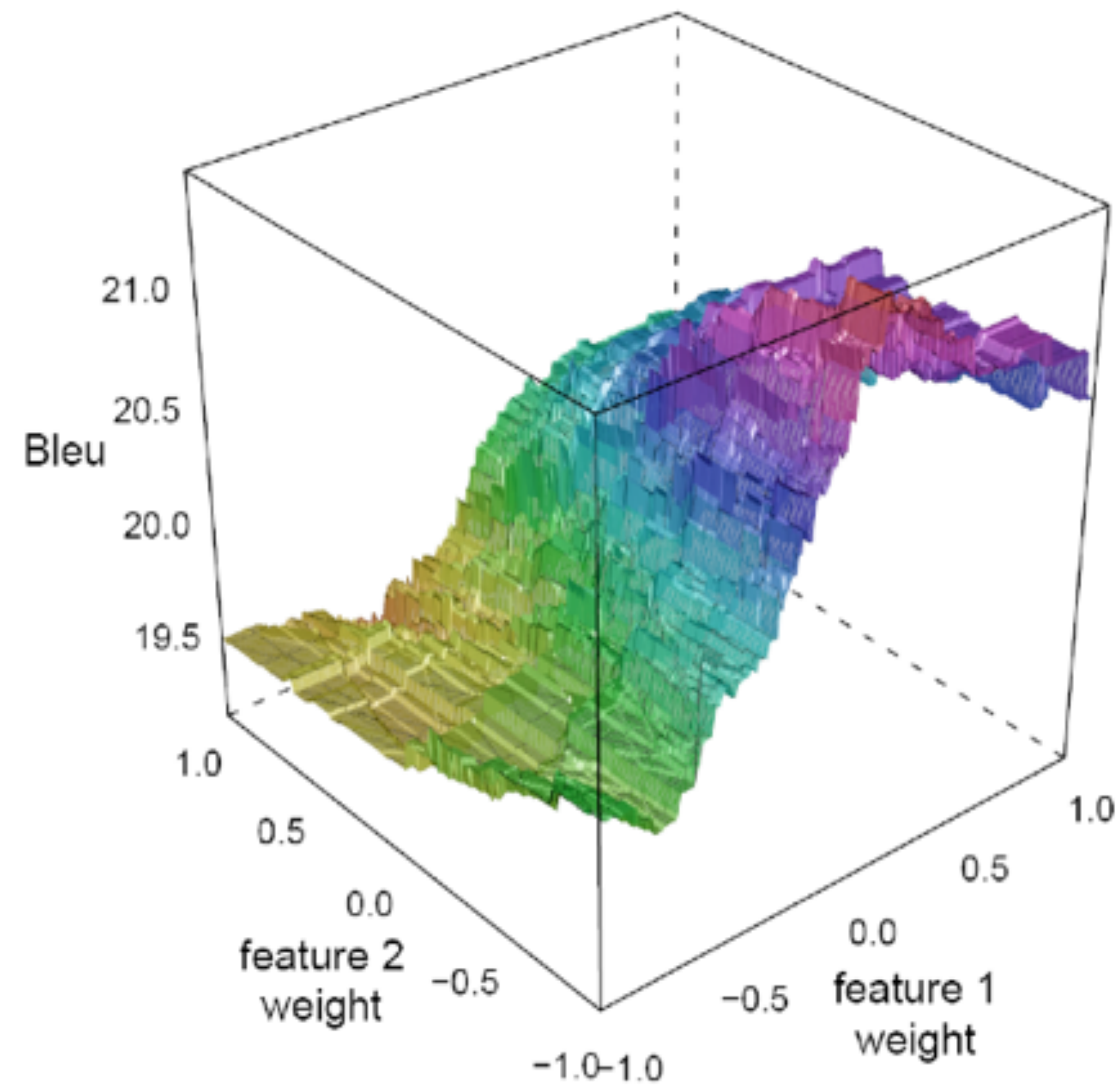
- Comparison of
  - *good statistical* system: high BLEU, high adequacy/fluency
  - *bad statistical* sys. (trained on less data): low BLEU, low adequacy/fluency
  - *Systran*: lowest BLEU score, but high adequacy/fluency

How Good are Automatic Metrics?

- Do n-gram methods like BLEU overly favor certain types of systems?
- Automatic metrics still useful
- During development of one system, a better BLEU indicates a better system
- Evaluating different systems has to depend on human judgement
- What are some other evaluation ideas?

Minimizing Error/Maximizing Bleu

- Adjust parameters to minimize error (*L*) when translating a training set
- Error as a function of parameters is
  - *nonconvex*: not guaranteed to find optimum
  - *piecewise constant*: slight changes in parameters might not change the output.
- Usual method: optimize one parameter at a time with linear programming



Generative/Discriminative Reunion

- Generative models can be cheap to train: “count and normalize” when nothing’s hidden.
- Discriminative models focus on problem: “get better translations”.
- Popular combination
  - Estimate several generative translation and language models using relative frequencies.
  - Find their optimal (log-linear) combination using discriminative techniques.

Generative/Discriminative Reunion

Score each hypothesis with several generative models:

$$\theta_1 p_{phrase}(\bar{s} \mid \bar{t}) + \theta_2 p_{phrase}(\bar{t} \mid \bar{s}) + \theta_3 p_{lexical}(s \mid t) + \ldots + \theta_7 p_{LM}(\bar{t}) + \theta_8 \#\,\mathrm{words} + \ldots$$

If necessary, renormalize into a probability distribution:

$$Z = \sum_k \exp(\boldsymbol{\theta} \cdot \mathbf{f}_k)$$

Unnecessary if thetas sum to 1 and p's are all probabilities.

where *k* ranges over all hypotheses. We then have

$$p(t_i \mid s) = \frac{1}{Z} \exp(\boldsymbol{\theta} \cdot \mathbf{f})$$

Exponentiation makes it positive.

for any given hypothesis *i*.

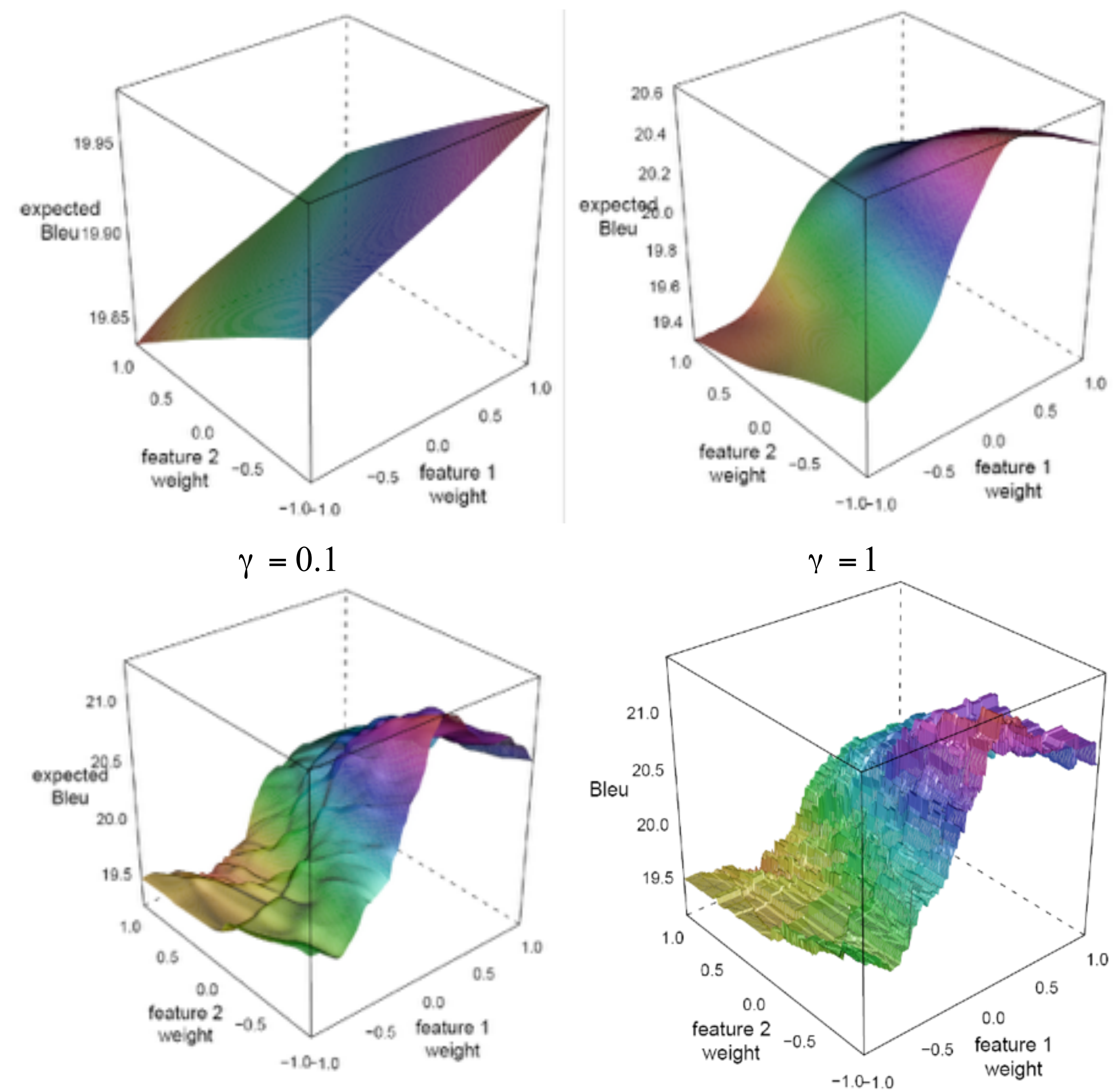
Minimizing Risk

Instead of the error of the 1-best translation, compute **expected error** (risk) using *k*-best translations; this makes the function differentiable.

Smooth probability estimates using gamma to even out local bumpiness. Gradually increase gamma to approach the 1-best error.

$$E_{p_{\gamma,\theta}}[L(s,t)]$$

$$p_{\gamma,\theta}(t_i \mid s_i) = \frac{[\exp \boldsymbol{\theta} \cdot \mathbf{f}_i]^\gamma}{\sum_{k'} [\exp \boldsymbol{\theta} \cdot \mathbf{f}_{k'}]^\gamma}$$



$\gamma = 0.1$ $\gamma = 1$ $\gamma = 10$ $\gamma = \infty$

Encoder-Decoder Models

(cf. Socher & Manning 2016)

Language Models

A language model computes a probability for a sequence of words: $P(w_1, \ldots, w_T)$

- Useful for machine translation
  - Word ordering:
    p(the cat is small) > p(small the is cat)
  - Word choice:
    p(walking home after school) > p(walking house after school)

Ngram LMs

- Performance improves with keeping around higher n-grams counts and doing smoothing and so-called backoff (e.g. if 4-gram not found, try 3-gram, etc)
- There are A LOT of n-grams!
  → Gigantic RAM requirements!
- Recent state of the art: *Scalable Modified Kneser-Ney Language Model Estimation* by Heafield et al.:
  "Using one machine with 140 GB RAM for 2.8 days, we built an unpruned model on 126 billion tokens"

Remember Word Embeddings

Standard Word Representation

The vast majority of rule-based **and** statistical NLP work regards words as atomic symbols: hotel, conference, walk

In vector space terms, this is a vector with one 1 and a lot of zeroes

[0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]

Dimensionality: 20K (speech) – 50K (PTB) – 500K (big vocab) – 13M (Google 1T)

We call this a "one-hot" representation. Its problem:

motel [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0] AND
hotel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0] = 0

Distributional Similarity

You can get a lot of value by representing a word by means of its neighbors

"You shall know a word by the company it keeps"

(J. R. Firth 1957: 11)

One of the most successful ideas of modern statistical NLP

government debt problems turning into banking crises as has happened in

saying that Europe needs unified banking regulation to replace the hodgepodge

↖ These words will represent *banking* ↗

You can vary whether you use local or large context to get a more syntactic or semantic clustering

Hard/Soft Clustering

Class based models learn word classes of similar words based on distributional information (~ class HMM)

- Brown clustering (Brown et al. 1992)
- Exchange clustering (Martin et al. 1998, Clark 2003)
- Desparsification and great example of unsupervised pre-training

Soft clustering models learn for each cluster/topic a distribution over words of how likely that word is in each cluster

- Latent Semantic Analysis (LSA/LSI), Random projections
- Latent Dirichlet Analysis (LDA), HMM clustering

Distributed Representation

Similar idea

Combine vector space semantics with the prediction of probabilistic models (Bengio et al. 2003, Collobert & Weston 2008, Turian et al. 2010)

In all of these approaches, including deep learning models, a word is represented as a dense vector

$$\textit{linguistics} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

Visualizing Embeddings

Vector Semantics

Mikolov, Yih & Zweig (2013)

These representations are *way* better at encoding dimensions of similarity than we realized!

- Analogies testing dimensions of similarity can be solved quite well just by doing vector subtraction in the embedding space

  Syntactically

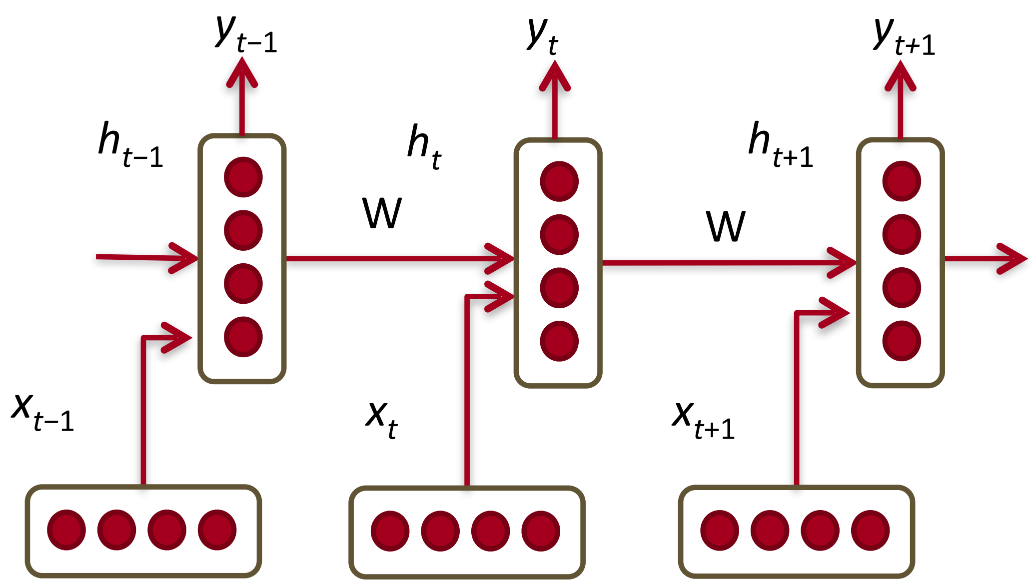
  - $x_{apple} - x_{apples} \approx x_{car} - x_{cars} \approx x_{family} - x_{families}$
  - Similarly for verb and adjective morphological forms

  Semantically (Semeval 2012 task 2)

  - $x_{shirt} - x_{clothing} \approx x_{chair} - x_{furniture}$

Recurrent Neural Nets

- RNNs tie the weights at each time step
- Condition the neural network on all previous words
- RAM requirement only scales with number of words



RNN LMs

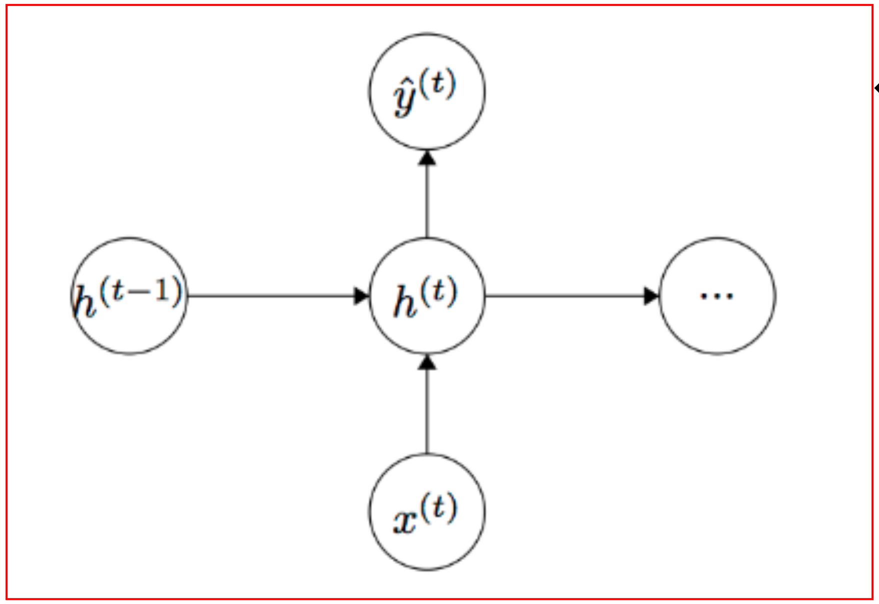
Given list of word **vectors**: $x_1, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_T$

At a single time step:

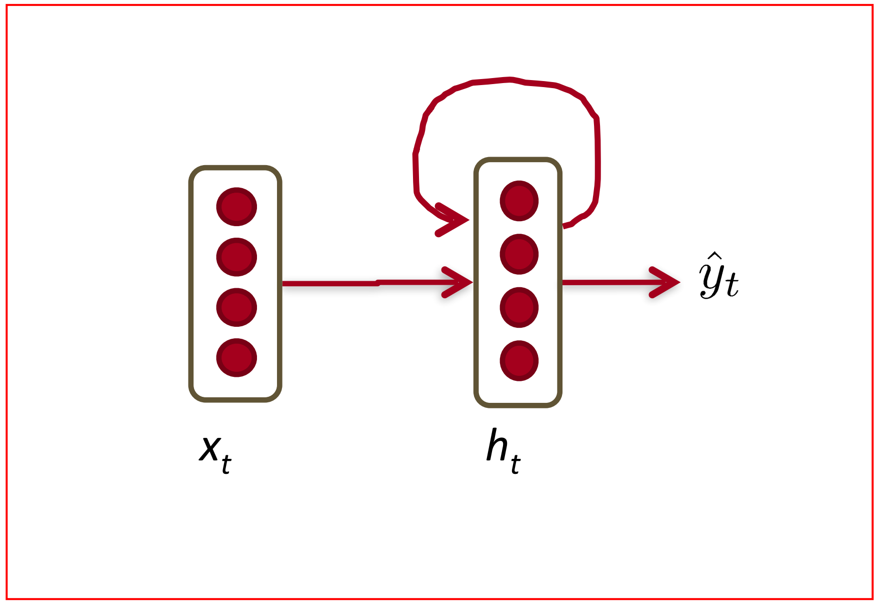
$$h_t = \sigma\left(W^{(hh)} h_{t-1} + W^{(hx)} x_{[t]}\right)$$

$$\hat{y}_t = \text{softmax}\left(W^{(S)} h_t\right)$$

$$\hat{P}(x_{t+1} = v_j \mid x_t, \ldots, x_1) = \hat{y}_{t,j}$$



←→



RNN LMs

Main idea: we use the same set of W weights at all time steps!

Everything else is the same:

$$
\begin{aligned}
h_t &= \sigma\left(W^{(hh)} h_{t-1} + W^{(hx)} x_{[t]}\right) \\
\hat{y}_t &= \mathrm{softmax}\left(W^{(S)} h_t\right) \\
\hat{P}(x_{t+1} = v_j \mid x_t, \ldots, x_1) &= \hat{y}_{t,j}
\end{aligned}
$$

$h_0 \in \mathbb{R}^{D_h}$ is some initialization vector for the hidden layer at time step 0

$x_{[t]}$ is the column vector of L at index [t] at time step t

$W^{(hh)} \in \mathbb{R}^{D_h \times D_h}$ $W^{(hx)} \in \mathbb{R}^{D_h \times d}$ $W^{(S)} \in \mathbb{R}^{|V| \times D_h}$

RNN LMs

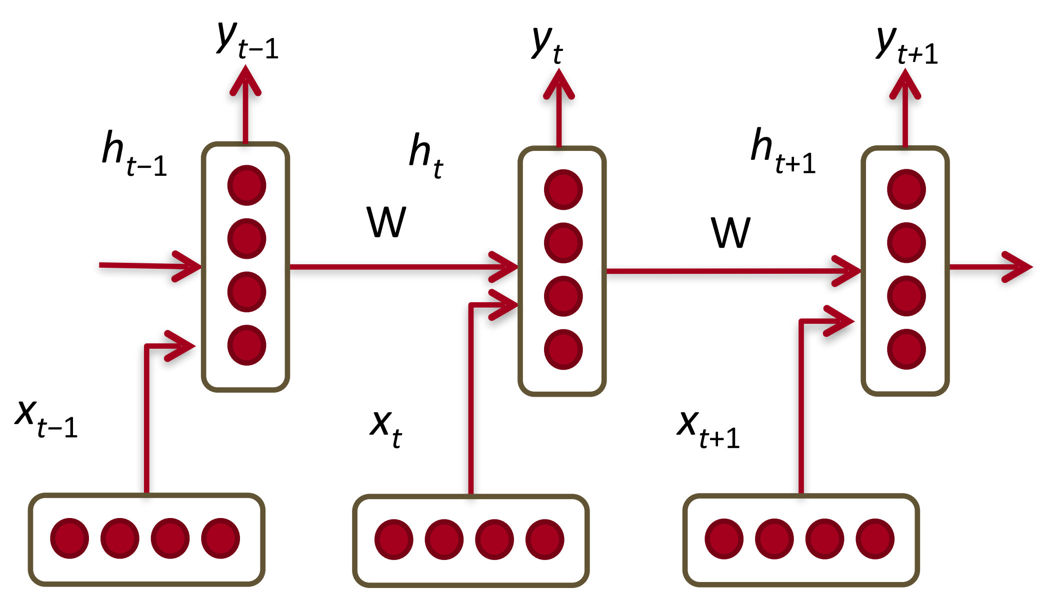
$\hat{y} \in \mathbb{R}^{|V|}$ is a probability distribution over the vocabulary

Same cross entropy loss function but predicting words instead of classes

$$J^{(t)}(\theta) = -\sum_{j=1}^{|V|} y_{t,j} \log \hat{y}_{t,j}$$

Training RNNs is hard!

- Multiply the same matrix at each time step during forward prop



- Ideally inputs from many time steps ago can modify output *y*
- Take $\frac{\partial E_2}{\partial W}$ for an example RNN with 2 time steps! Insightful!

Clipping Gradients

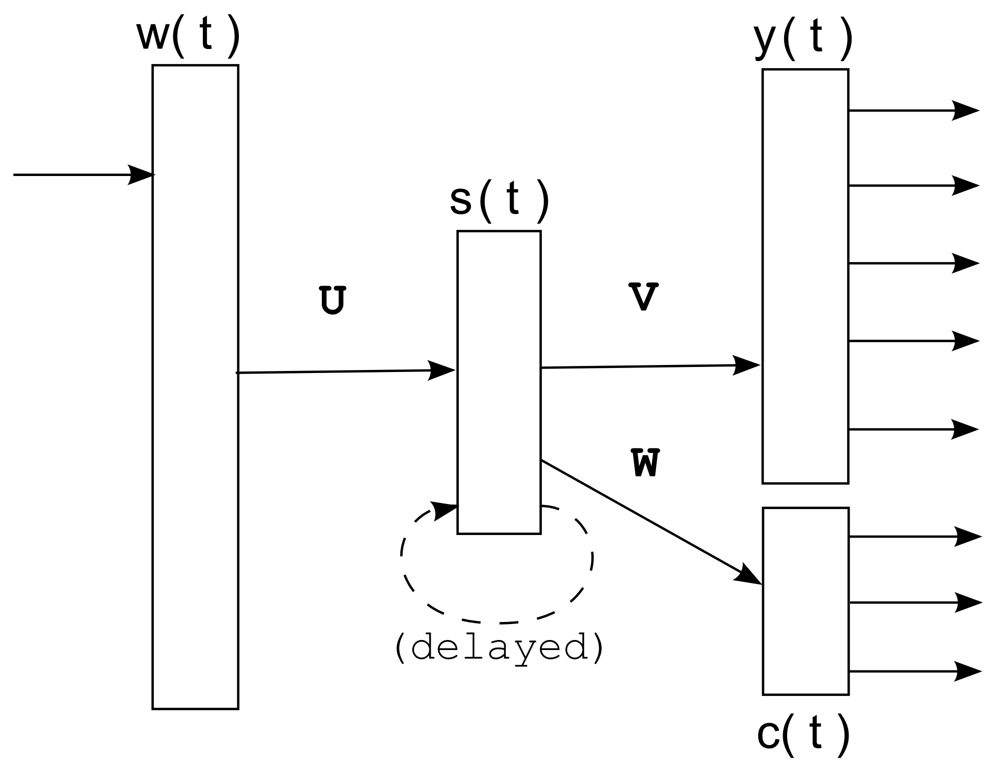
- The solution first introduced by Mikolov is to clip gradients to a maximum value.

Algorithm 1 Pseudo-code for norm clipping the gradients whenever they explode

$\hat{\mathbf{g}} \leftarrow \frac{\partial \mathcal{E}}{\partial \theta}$
if $\|\hat{\mathbf{g}}\| \geq threshold$ **then**
  $\hat{\mathbf{g}} \leftarrow \frac{threshold}{\|\hat{\mathbf{g}}\|} \hat{\mathbf{g}}$
end if

- Makes a big difference in RNNs.

Slow Softmax? Class Layer



$$P(w_i|history) = P(c_i|\mathbf{s}(t))P(w_i|c_i,\mathbf{s}(t)) \quad (1)$$

- Words are assigned to ”classes” based on their unigram frequency
- First, class layer is evaluated; then, only words belonging to the predicted class are evaluated, instead of the whole output layer $\mathbf{y}$ [Goodman2001]
- Provides speedup in some cases more than $100\times$

Perplexity Results

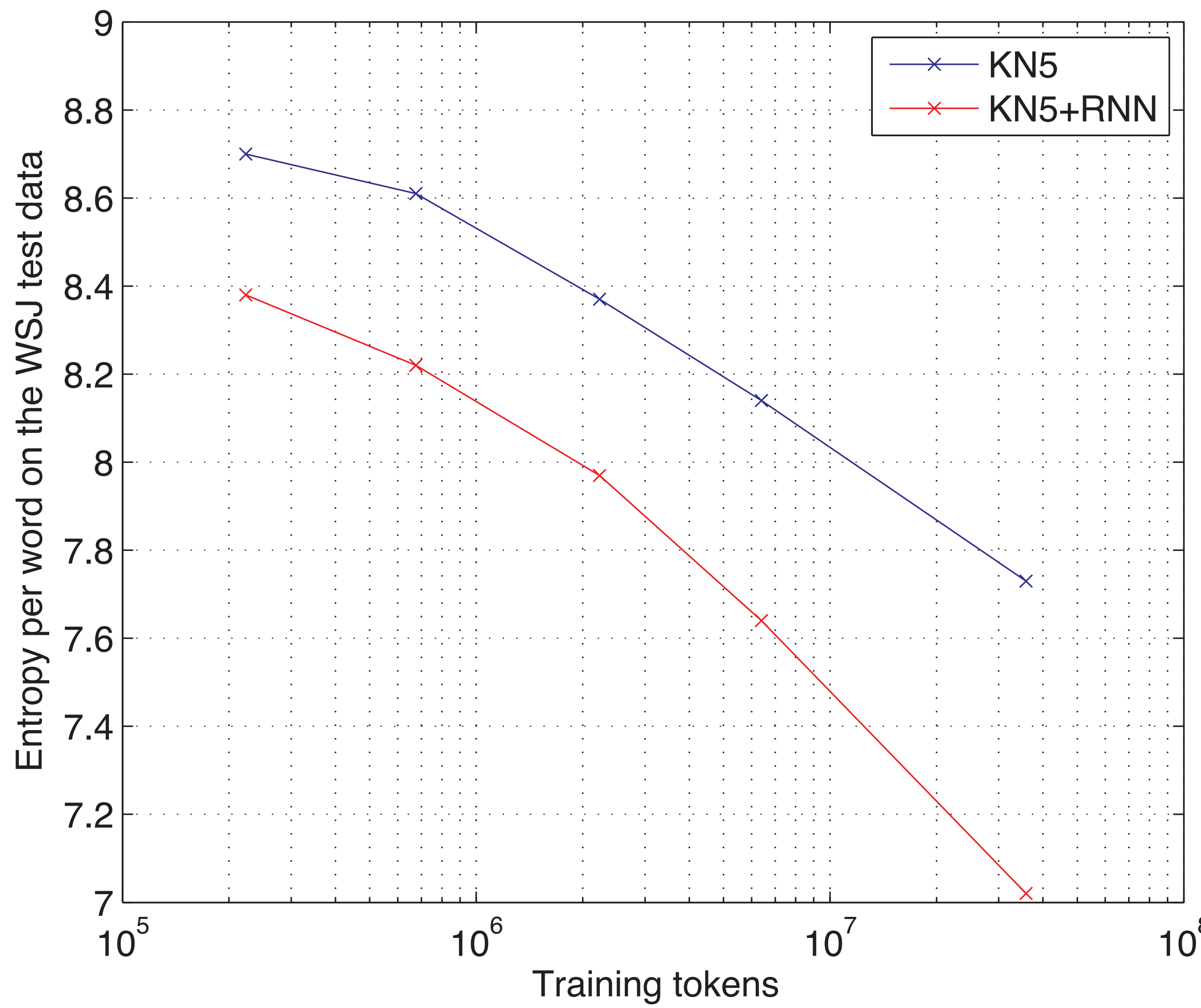
KN5 = Count-based language model with Kneser-Ney smoothing & 5-grams

Table 2. *Comparison of different neural network architectures on Penn Corpus (1M words) and Switchboard (4M words).*

	Penn Corpus		Switchboard	
Model	NN	NN+KN	NN	NN+KN
KN5 (baseline)	-	141	-	92.9
feedforward NN	141	118	85.1	77.5
RNN trained by BP	137	113	81.3	75.4
RNN trained by BPTT	123	106	77.5	72.5

Table from paper *Extensions of recurrent neural network language model* by Mikolov et al 2011

Learning Curves

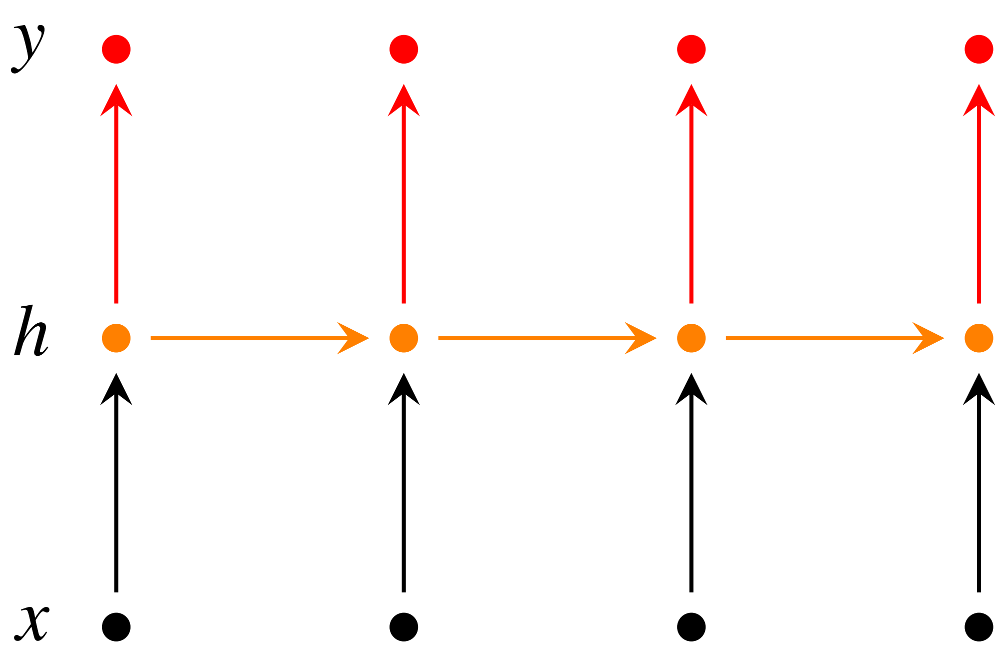


- The improvement obtained from a single RNN model over the best backoff model **increases** with more data!

Deeper

(Irsoy & Cardie, 2014)

RNNs



$$h_t = f(Wx_t + Vh_{t-1} + b)$$
$$y_t = g(Uh_t + c)$$

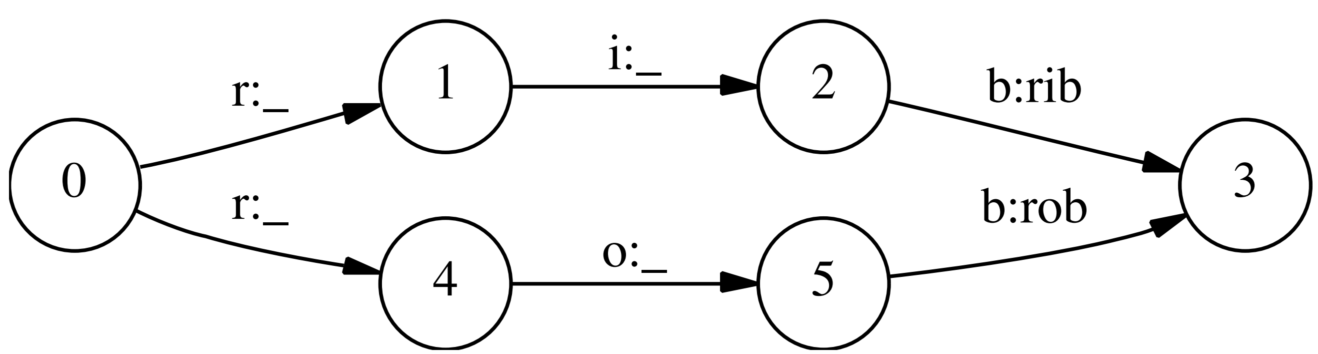
x represents a token (word) as a vector.

y represents the output label.

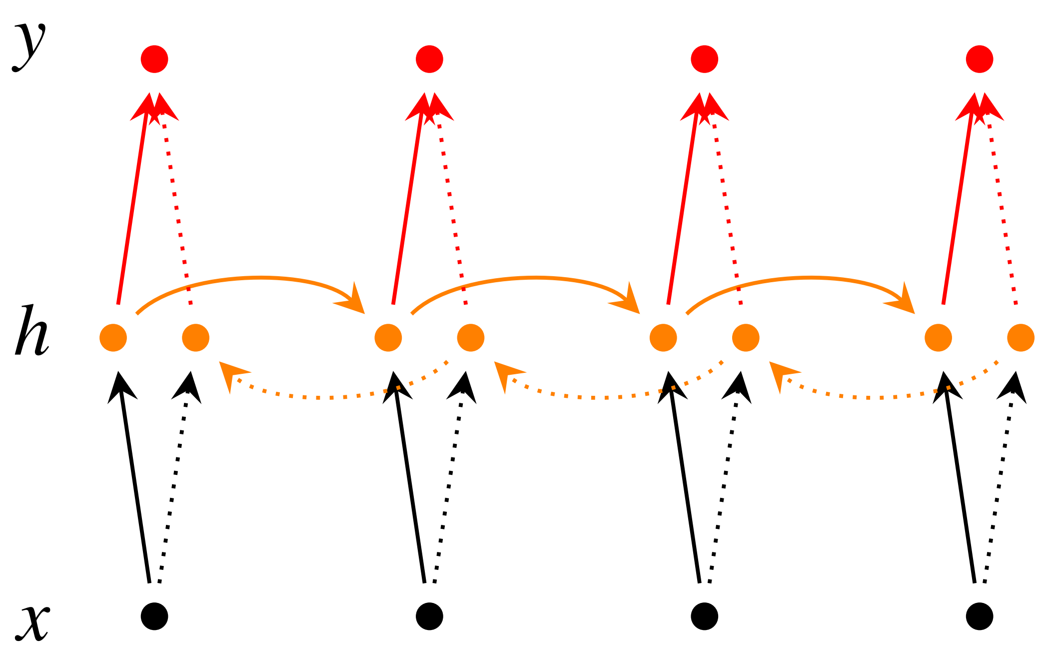
h is the memory, computed from the past memory and current word. It summarizes the sentence up to that time.

Label Bias

- In some state space configurations, MEMMs (and RNNs) essentially ignore the inputs
- This is not a problem for HMMs and CRFs



Bidirectionality



$$\overrightarrow{h}_t = f(\overrightarrow{W}x_t + \overrightarrow{V}\overrightarrow{h}_{t-1} + \overrightarrow{b})$$

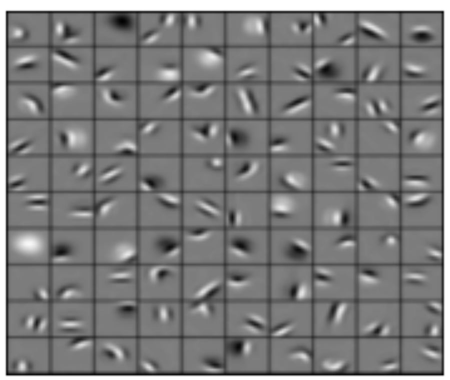
$$\overleftarrow{h}_t = f(\overleftarrow{W}x_t + \overleftarrow{V}\overleftarrow{h}_{t+1} + \overleftarrow{b})$$

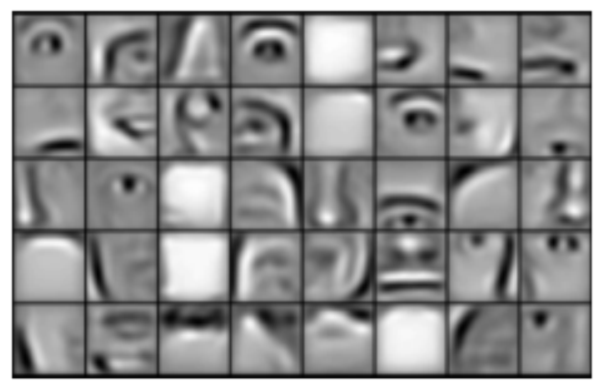
$$y_t = g(U[\overrightarrow{h}_t;\overleftarrow{h}_t] + c)$$

$h = [\overrightarrow{h};\overleftarrow{h}]$ now represents (summarizes) the past and future around a single token.

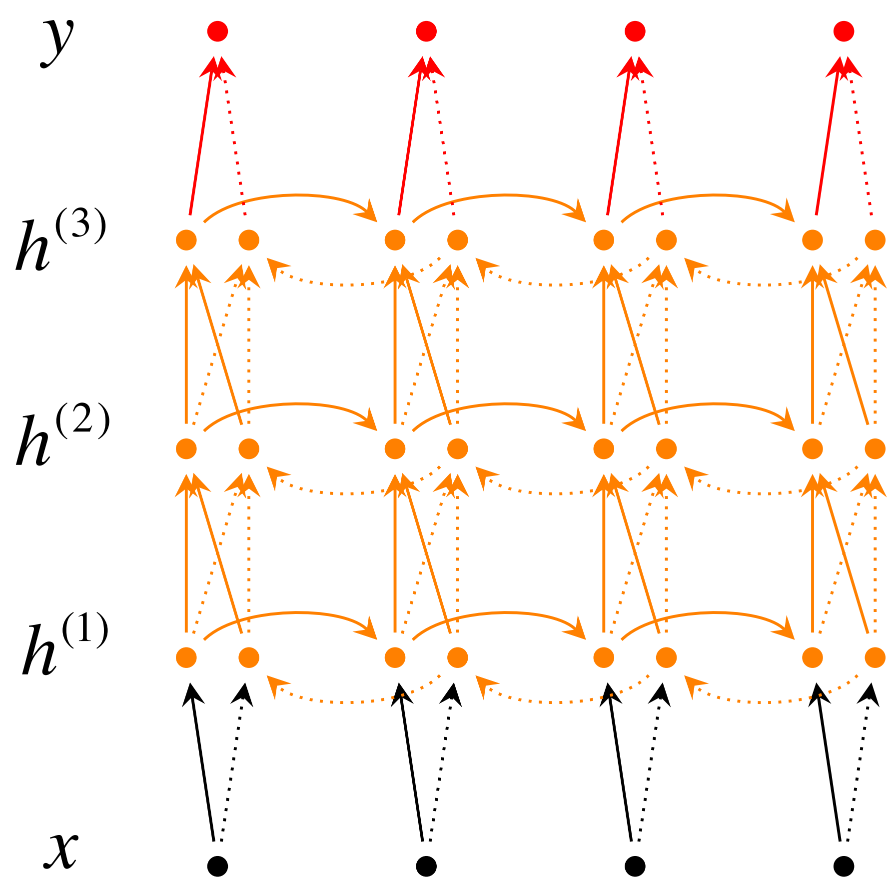
Going Deep

Are recurrent networks really *deep*? (e.g. like this)





Going Deep



$$\overrightarrow{h}_t^{(i)} = f(\overrightarrow{W}^{(i)} h_t^{(i-1)} + \overrightarrow{V}^{(i)} \overrightarrow{h}_{t-1}^{(i)} + \overrightarrow{b}^{(i)})$$

$$\overleftarrow{h}_t^{(i)} = f(\overleftarrow{W}^{(i)} h_t^{(i-1)} + \overleftarrow{V}^{(i)} \overleftarrow{h}_{t+1}^{(i)} + \overleftarrow{b}^{(i)})$$

$$y_t = g(U[\overrightarrow{h}_t^{(L)}; \overleftarrow{h}_t^{(L)}] + c)$$

Each memory layer passes an intermediate sequential representation to the next.

Opinion Mining

Fine-grained opinion analysis aims to detect subjectivity (e.g. “hate”) and characterize

- Intensity (e.g. strong)
- Sentiment (e.g. negative)
- Opinion holder, target or topic
- …

Important for a variety of NLP tasks such as

- Opinion-oriented question answering
- Opinion summarization

Opinion Mining

In this work, we focus on detecting
direct subjective expressions (DSEs) and
expressive subjective expressions (ESEs).

DSE: Explicit mentions of private states or speech events expressing private states

ESE: Expressions that indicate sentiment, emotion, etc. without explicitly conveying them.

Example

The committee, [as usual]$_{ESE}$, [has refused to make any statements]$_{DSE}$.

In BIO notation (where a token is the atomic unit):

The	committee	,	as	usual	,	has
O	O	O	B_ESE	I_ESE	O	B_DSE

refused	to	make	any	statements	.
I_DSE	I_DSE	I_DSE	I_DSE	I_DSE	O

CRF et al.

Success of CRF based approaches hinges critically on access to a good feature set, typically based on

- Constituency and dependency parse trees
- Manually crafted opinion lexicons
- Named entity taggers
- Other preprocessing components

(See Yang and Cardie (2012) for an up-to-date list.)

What about feature learning?

Hypotheses

We expected that deep recurrent nets would improve upon shallow recurrent nets, especially on ESE extraction.

- ESEs are harder to identify: They are variable in length and might involve terms that are neutral in most contexts (e.g. “as usual”, “in fact”).

How the networks would perform against (semi)CRFs was unclear, especially when CRFs are given access to word vectors.

Results: Examples

True	The situation obviously remains fluid from hour to hour but it [seems to be] [going in the right direction]
Deep RntNN	The situation [obviously] remains fluid from hour to hour but it [seems to be going in the right] direction
Shallow RntNN	The situation [obviously] remains fluid from hour to hour but it [seems to be going in] the right direction
Semi-CRF	The situation [obviously remains fluid from hour to hour but it seems to be going in the right direction]