Statistical Machine Translation
Part VI – Better Word Alignment, Morphology and Syntax

Alexander Fraser

CIS, LMU München

2016.12.20   SMT and NMT

MT talk on January 10th

- Christine Bruckner will give a talk and demo:

  Machine Translation and the Professional Translator's Workplace – Practical Insights into Current Commercial Solutions

- Talk on Tues. January 10th at 12:15 in 131 (upstairs, near CIS)

Back to SMT

- We changed the seminar schedule
  - I will actually go back to SMT in this lecture
  - I'm going to talk about some other areas of importance in SMT research
  - Touches on work in my research group
- This lecture was originally designed to be after the last SMT lecture
- But I'll try to make very general comments about problems in NMT as appropriate
- Matthias Huck will present the details of how NMT works in January

Where we have been

- We've discussed the MT problem and evaluation
- We have covered phrase-based SMT
  - Model (now using log-linear model)
  - Training of phrase block distribution
    - Dependent on word alignment
  - Search
  - Evaluation

Where we are going

- Word alignment makes linguistic assumptions that are not realistic
- Phrase-based decoding makes linguistic assumptions that are not realistic
- How can we improve on this?

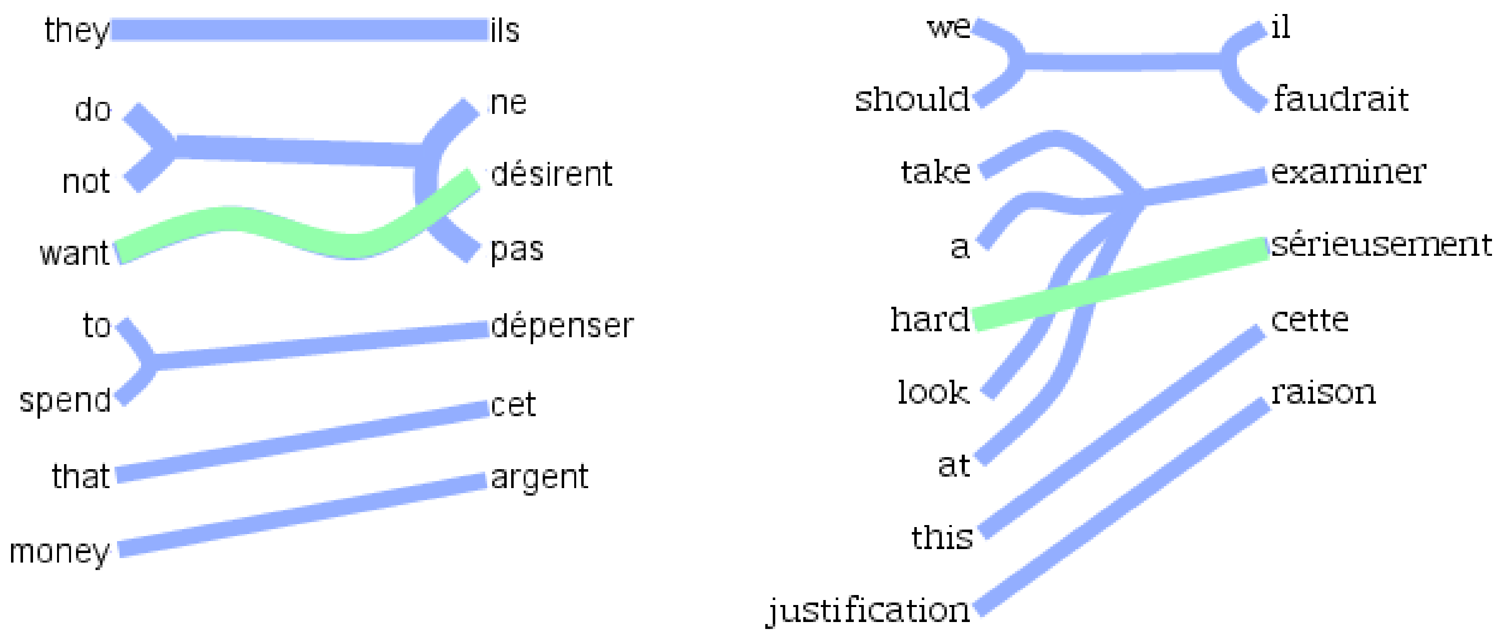
Outline

- Improved word alignment
- Morphology
- Syntax
- Conclusion

Improved word alignments

- My dissertation was on word alignment
- Three main pieces of work
  - Measuring alignment quality (F-alpha)
    - We saw this already
  - A new generative model with many-to-many structure
  - A hybrid discriminative/generative training technique for word alignment

Modeling the Right Structure



- 1-to-N assumption
  - Multi-word “cepts” (words in one language translated as a unit) only allowed on target side. Source side limited to single word “cepts”.
- Phrase-based assumption
  - “cepts” must be consecutive words

LEAF Generative Story

source	absolutely	[comma]	they	do	not	want	to	spend	that	money	
word type (1)	DEL.	DEL.	HEAD	non-head	HEAD	HEAD	non-head	HEAD	HEAD	HEAD	
linked from (2)			THEY	do	NOT	WANT	to	SPEND	THAT	MONEY	
head(3)			ILS		PAS	DESIRENT		DEPENSER	CET	ARGENT	
cept size(4)			1		2	1		1	1	1	
num spurious(5)	1										
spurious(6)	aujourd'hui										
non-head(7)			ILS	PAS	ne	DESIRENT		DEPENSER	CET	ARGENT	
placement(8)	aujourd'hui		ILS	ne	DESIRENT	PAS		DEPENSER	CET	ARGENT	
spur. placement(9)			ILS	ne	DESIRENT	PAS		DEPENSER	CET	ARGENT	aujourd'hui

- Explicitly model three word types:
  - **Head word**: provide most of conditioning for translation
    - Robust representation of multi-word cepts (for this task)
    - This is to semantics as ``syntactic head word'' is to syntax
  - **Non-head word**: attached to a head word
  - **Deleted source words** and **spurious target words** (NULL aligned)

LEAF Generative Story

source	absolutely	[comma]	they	do	not	want	to	spend	that	money	
word type (1)	DEL.	DEL.	HEAD	non-head	HEAD	HEAD	non-head	HEAD	HEAD	HEAD	
linked from (2)			THEY	do	NOT	WANT	to	SPEND	THAT	MONEY	
head(3)			ILS		PAS	DESIRENT		DEPENSER	CET	ARGENT	
cept size(4)			1		2	1		1	1	1	
num spurious(5)	1										
spurious(6)	aujourd'hui										
non-head(7)			ILS	PAS	ne	DESIRENT		DEPENSER	CET	ARGENT	
placement(8)	aujourd'hui		ILS	ne	DESIRENT	PAS		DEPENSER	CET	ARGENT	
spur. placement(9)			ILS	ne	DESIRENT	PAS		DEPENSER	CET	ARGENT	aujourd'hui

- Once source cepts are determined, exactly one target head word is generated from each source head word
- Subsequent generation steps are then conditioned on a single target and/or source head word
- See EMNLP 2007 paper for details

Discussion

- LEAF is a powerful model
- But, exact inference is intractable
  - We use hillclimbing search from an initial alignment
- Models correct structure: M-to-N discontiguous
  - First general purpose statistical word alignment model of this structure!
    - Can get 2nd best, 3rd best, etc hypothesized alignments (unlike 1-to-N models combined with heuristics)
  - Head word assumption allows use of multi-word cepts
    - Decisions robustly decompose over words (not phrases)

New knowledge sources for word alignment

- It is difficult to add new knowledge sources to generative models
  - Requires completely reengineering the generative story for each new source
- Existing unsupervised alignment techniques can not use manually annotated data

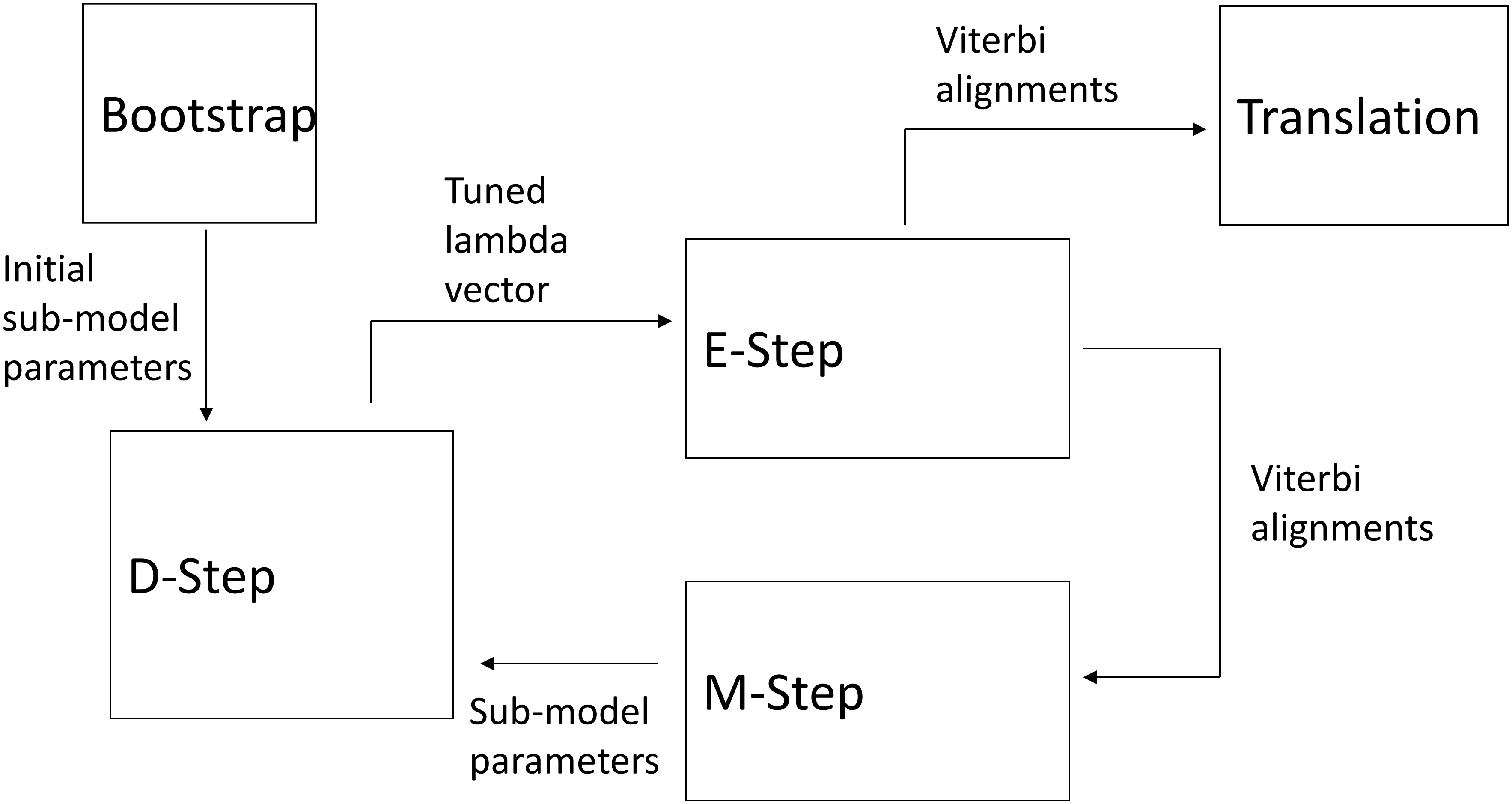
Decomposing LEAF

- Decompose each step of the LEAF generative story into a sub-model of a log-linear model
  - Add backed off forms of LEAF sub-models
  - Add heuristic sub-models (do not need to be related to generative story!)
  - Allows tuning of vector λ which has a scalar for each sub-model controlling its contribution
- How to train this log-linear model?

Semi-Supervised Training

- Define a semi-supervised algorithm which alternates increasing likelihood with decreasing error
  - Increasing likelihood is similar to EM
  - Discriminatively bias EM to converge to a local maxima of likelihood which corresponds to “better” alignments
    - “Better” = higher F_{α}-score on small gold standard word alignments corpus
    - Integrate minimization from MERT together with EM

The EMD Algorithm



Discussion

- Usual formulation of semi-supervised learning: "using unlabeled data to help supervised learning"
  - Build initial supervised system using labeled data, predict on unlabeled data, then iterate
  - But we do not have enough gold standard word alignments to estimate parameters directly!
- EMD allows us to train a small number of important parameters discriminatively, the rest using likelihood maximization, and allows interaction
  - Similar in spirit (but not details) to semi-supervised clustering

Contributions

- Found a metric for measuring alignment quality which correlates with decoding quality
- Designed LEAF, the first generative model of M-to-N discontiguous alignments
- Developed a semi-supervised training algorithm, the EMD algorithm
  - Allows easy incorporation of new features into a word alignment model that is still mostly unsupervised
- Obtained large gains of 1.2 BLEU and 2.8 BLEU points for French/English and Arabic/English tasks

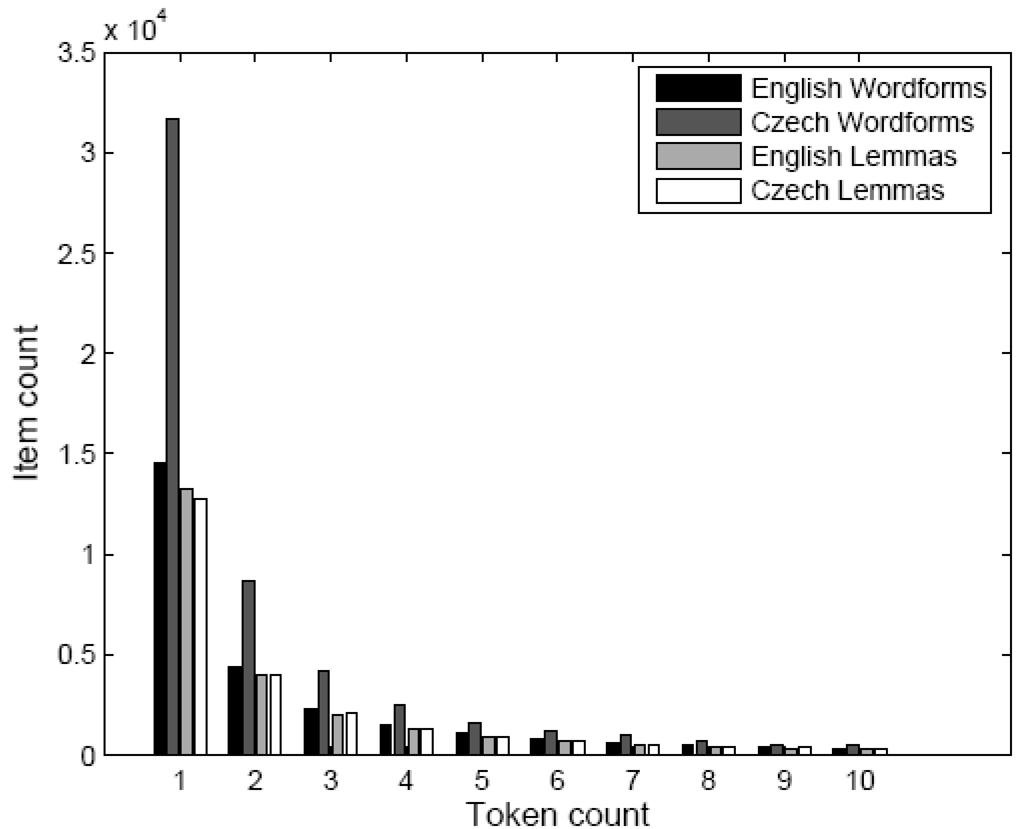
Outlook

- There was a lot of interest in word alignment around 2005-2009
  - Key to phrase-based approach – need good quality word alignments, particularly for sparsely seen vocabulary
  - Word alignment is still useful for many specialized subproblems in translation and related multilingual problems
- However, neural machine translation is not trained on word alignments!
  - As a side effect of training on sentence pairs, a so-called "attentional model" is learned
  - Gives weight to the input embeddings of words that will be useful for translating the current word being generated
- However, ideas from word alignment are still being integrated into the neural model, this will probably continue for a few years

Morphology

- We will use the term morphology loosely here
  - We will discus two main phenomena: Inflection, Compounding
  - There is less work in SMT on modeling of these phenomena than there is on syntactic modeling
    - A lot of work on morphological reduction (e.g., make it like English if the target language is English)
    - Not much work on generating (necessary to translate to, for instance, Slavic languages or Finnish)

Inflection



Goldwater and McClosky 2005

Inflection

- Inflection
  - The best ideas here are to strip redundant morphology
    - For instance case markings that are not used in target language
  - Can also add pseudo-words
    - One interesting paper looks at translating Czech to English (Goldwater and McClosky)
    - Inflection which should be translated to a pronoun is simply replaced by a pseudo-word to match the pronoun in preprocessing

Compounds

- Find the best split by using word frequencies of components (Koehn 2003)
- Aktionsplan -> Akt Ion Plan or   Aktion Plan?
  - Since Ion (English: ion) is not frequent, do not pick such a splitting!
- Initially not improved by using hand-crafted morphological knowledge
- Fabienne Cap has shown using SMOR (Stuttgart Morphological Analyzer) together with corpus statistics is better (Fritzinger and Fraser WMT 2010)

Work at Munich on Morphology

- My group has done a lot of work on modeling inflection and compounds in SMT
  - Particularly for translation into morphologically rich languages (e.g., English to German translation)
- Looking at applying similar techniques in NMT

Syntax

- Better modeling of syntax was a very hot topic in SMT
- For instance, consider the problem of translating German to English
  - One way to deal with this is to make German look more like English

Clause Level Restructuring [Collins et al.]

- Why **clause structure**?
  - languages *differ vastly* in their clause structure (English: SVO, Arabic: VSO, German: fairly *free order*; a lot details differ: position of adverbs, sub clauses, etc.)
  - large-scale restructuring is a *problem* for phrase models
- **Restructuring**
  - *reordering* of constituents (main focus)
  - add/drop/change of *function words*

Slide from Koehn and Lopez 2008

Clause Structure

```
S   PPER-SB  Ich     I
    VAFIN-HD werde    will
    VP-OC    PPER-DA Ihnen   you
             NP-OA   ART-OA  die   the
                     ADJ-NK  entsprechenden    corresponding
                     NN-NK   Anmerkungen   comments
             VVFIN   aushaendigen     pass on
             $,      ,    ,
             S-MO    KOUS-CP damit   so that
                     PPER-SB Sie    you
                     VP-OC   PDS-OA  das   that
                             ADJD-MO eventuell    perhaps
                             PP-MO   APRD-MO  bei   in
                                     ART-DA   der   the
                                     NN-NK    Abstimmung  vote
                             VVINF   uebernehmen    include
                     VMFIN   koennen   can
$.  .             .
```

MAIN CLAUSE

SUB-ORDINATE CLAUSE

- *Syntax tree* from German parser

Slide from Koehn and Lopez 2008

Reordering When Translating

```
S     PPER-SB   Ich                              I
      VAFIN-HD  werde                            will
      PPER-DA   Ihnen                            you
      NP-OA     ART-OA  die                      the
                ADJ-NK  entsprechenden            corresponding
                NN-NK   Anmerkungen               comments
      VVFIN     aushaendigen                     pass on
$,    ,                                          ,
S-MO  KOUS-CP   damit                            so that
      PPER-SB   Sie                              you
      PDS-OA    das                              that
      ADJD-MO   eventuell                        perhaps
      PP-MO     APRD-MO  bei                     in
                ART-DA   der                      the
                NN-NK    Abstimmung               vote
      VVINF     uebernehmen                      include
      VMFIN     koennen                          can
$. .                                             .
```

- *Reordering* when translating into English
  - tree is *flattened*
  - clause level constituents line up

Slide from Koehn and Lopez 2008

Systematic Reordering German $\rightarrow$ English

- Many types of reorderings are **systematic**
  - *move verb group together*
  - *subject - verb - object*
  - *move negation in front of verb*

$\Rightarrow$ *Write rules by hand*

- apply rules to test and training data
- train standard *phrase-based* SMT system

Slide from Koehn and Lopez 2008

English to German

- A lot of work in Munich on this language pair
- We can also apply this idea in translation from English to German
  - Put English in German word order
  - A bit more difficult but doable (Gojun and Fraser 2012)
    - More recent work also looks at agreement and tense

But what if we want to integrate probabilities?

- It turns out that we can!
- We will use something called a synchronous context free grammar (SCFG)
- This is surprisingly simple
  - Just involves defining a CFG with some markup showing what do to with the target language
  - We’ll first do a short example translating an English NP to a Chinese NP
  - Then we'll look at some German to English phenomena

Tree-Based Models

- Traditional statistical models operate on sequences of words

- Many translation problems can be best explained by pointing to syntax
  - reordering, e.g., verb movement in German–English translation
  - long distance agreement (e.g., subject-verb) in output

⇒ Translation models based on tree representation of language
  - significant ongoing research
  - state-of-the art for some language pairs

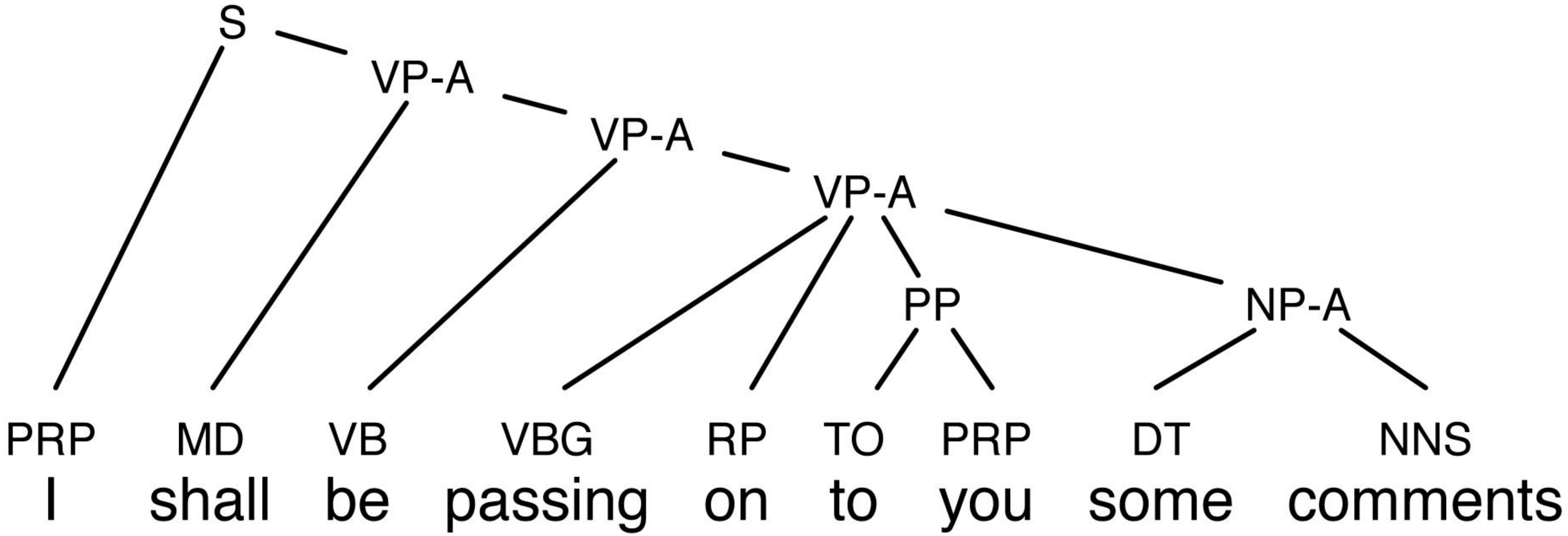
Slide from Koehn 2009

Phrase Structure Grammar

- Phrase structure
  - noun phrases: the big man, a house, ...
  - prepositional phrases: at 5 o'clock, in Edinburgh, ...
  - verb phrases: going out of business, eat chicken, ...
  - adjective phrases, ...

- Context-free Grammars (CFG)
  - non-terminal symbols: phrase structure labels, part-of-speech tags
  - terminal symbols: words
  - production rules: $\mathrm{NT} \rightarrow [\mathrm{NT},\mathrm{T}]+$
    example: $\mathrm{NP} \rightarrow \mathrm{DET}\ \mathrm{NN}$

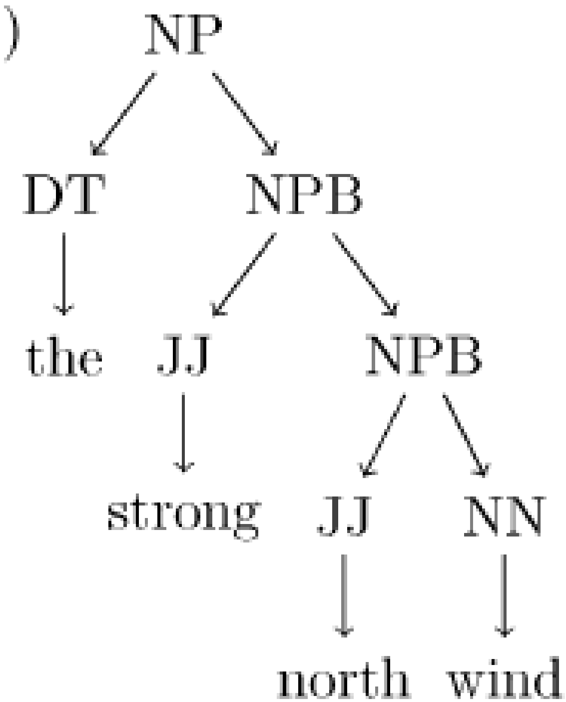
Phrase Structure Grammar



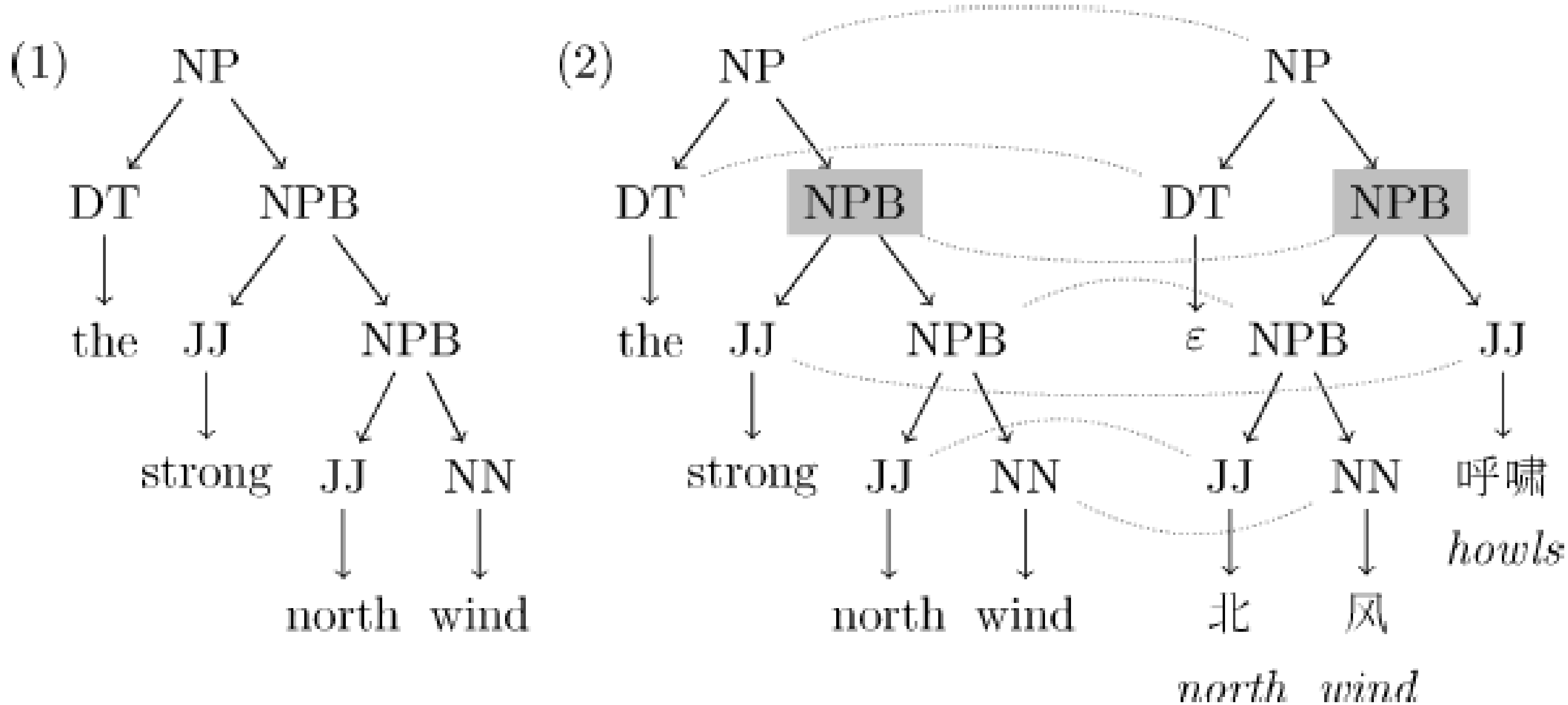
Phrase structure grammar tree for an English sentence
(as produced Collins' parser)

$$
\begin{array}{rcl}
\mathrm{NP} & \longrightarrow & \mathrm{DT\ NPB} \\
\mathrm{NPB} & \longrightarrow & \mathrm{JJ\ NPB} \\
\mathrm{NPB} & \longrightarrow & \mathrm{NP} \\
\mathrm{DT} & \longrightarrow & \text{the} \\
\mathrm{JJ} & \longrightarrow & \text{strong} \\
\mathrm{JJ} & \longrightarrow & \text{north} \\
\mathrm{NN} & \longrightarrow & \text{wind}
\end{array}
$$

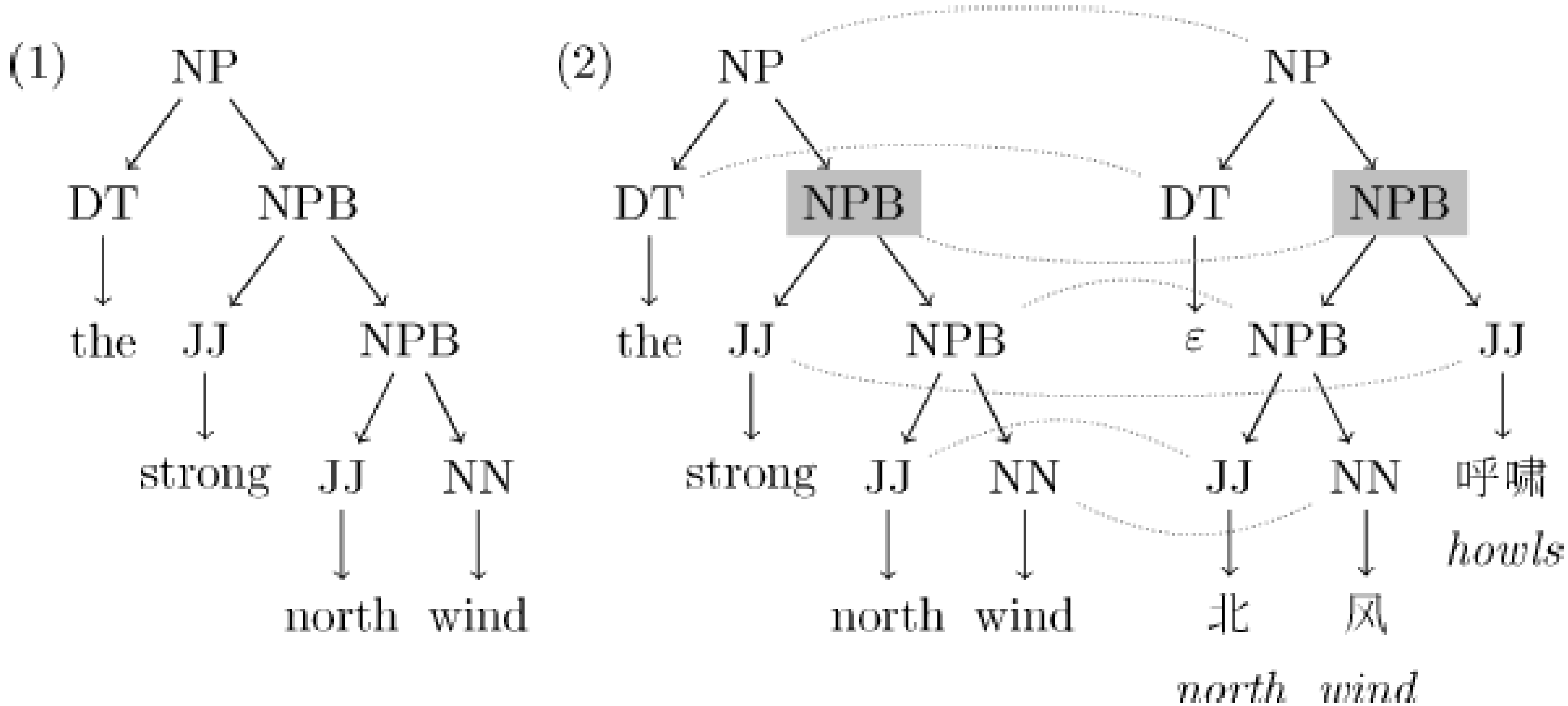
(1)



Lopez 2008


(1)
NP
DT
NPB
the
JJ
NPB
strong
JJ
NN
north
wind
(2)
NP
DT
NPB
the
JJ
NPB
strong
JJ
NN
north
wind
NP
DT
NPB
ε
NPB
JJ
JJ
NN
呼啸
howls
北
风
north
wind

Lopez 2008



$$\mathrm{NP} \longrightarrow \mathrm{DT}_{\boxed{1}}\mathrm{NPB}_{\boxed{2}} \ / \ \mathrm{DT}_{\boxed{1}}\mathrm{NPB}_{\boxed{2}}$$

$$\mathrm{NPB} \longrightarrow \mathrm{JJ}_{\boxed{1}}\mathrm{NN}_{\boxed{2}} \ / \ \mathrm{JJ}_{\boxed{1}}\mathrm{NN}_{\boxed{2}}$$

$$\mathrm{NPB} \longrightarrow \mathrm{NPB}_{\boxed{1}}\mathrm{JJ}_{\boxed{2}} \ / \ \mathrm{JJ}_{\boxed{2}}\mathrm{NPB}_{\boxed{1}}$$

$$\mathrm{DT} \longrightarrow \text{the} \ / \ \varepsilon$$

$$\mathrm{JJ} \longrightarrow \text{strong} \ / \ \text{呼啸}$$

$$\mathrm{JJ} \longrightarrow \text{north} \ / \ \text{北}$$

$$\mathrm{NN} \longrightarrow \text{wind} \ / \ \text{风}$$

Lopez 2008

Learning a SCFG from data

- We can learn rules of this kind
  - Given: Chinese/English parallel text
  - We parse the Chinese (so we need a good Chinese parser)
  - We parse the English (so we need a good English parser)
  - Then we word align the parallel text
  - Then we extract the aligned tree nodes to get SCFG rules; we can use counts to get probabilities

Synchronous Phrase Structure Grammar

- English rule

$$\text{NP} \longrightarrow \text{DET JJ NN}$$

- French rule

$$\text{NP} \longrightarrow \text{DET NN JJ}$$

- Synchronous rule (indices indicate alignment):

$$\text{NP} \longrightarrow \text{DET}_1 \ \text{NN}_2 \ \text{JJ}_3 \ \Big| \ \text{DET}_1 \ \text{JJ}_3 \ \text{NN}_2$$

Synchronous Grammar Rules

- Nonterminal rules

$$\text{NP} \rightarrow \text{DET}_1\ \text{NN}_2\ \text{JJ}_3 \mid \text{DET}_1\ \text{JJ}_3\ \text{NN}_2$$

- Terminal rules

$$\text{N} \rightarrow \text{maison} \mid \text{house}$$

$$\text{NP} \rightarrow \text{la maison bleue} \mid \text{the blue house}$$

- Mixed rules

$$\text{NP} \rightarrow \text{la maison JJ}_1 \mid \text{the JJ}_1\ \text{house}$$

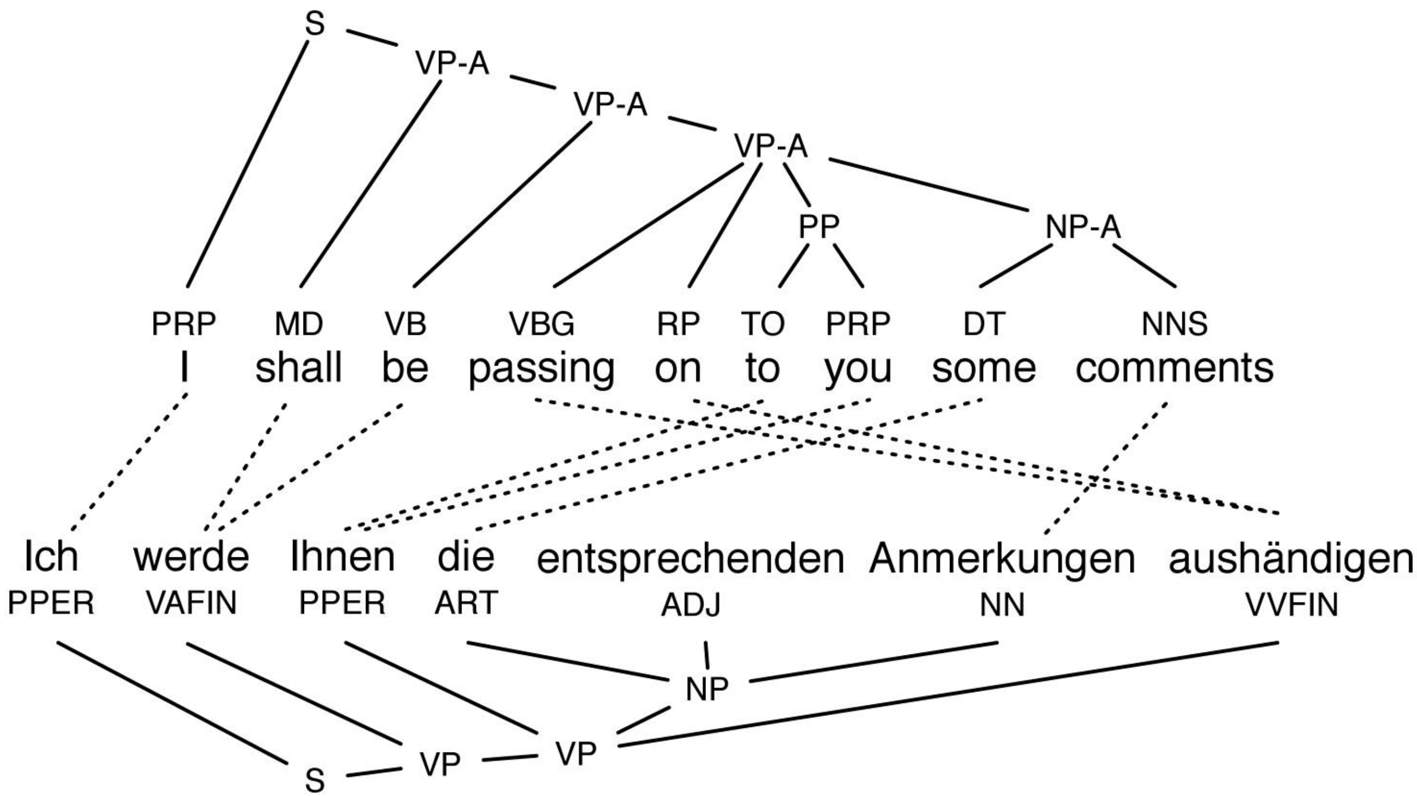
Tree-Based Translation Model

- Translation by parsing
  - synchronous grammar has to parse entire input sentence
  - output tree is generated at the same time
  - process is broken up into a number of rule applications
- Translation probability

$$\text{SCORE}(\text{TREE}, \text{E}, \text{F}) = \prod_i \text{RULE}_i$$

- Many ways to assign probabilities to rules

Aligned Tree Pair

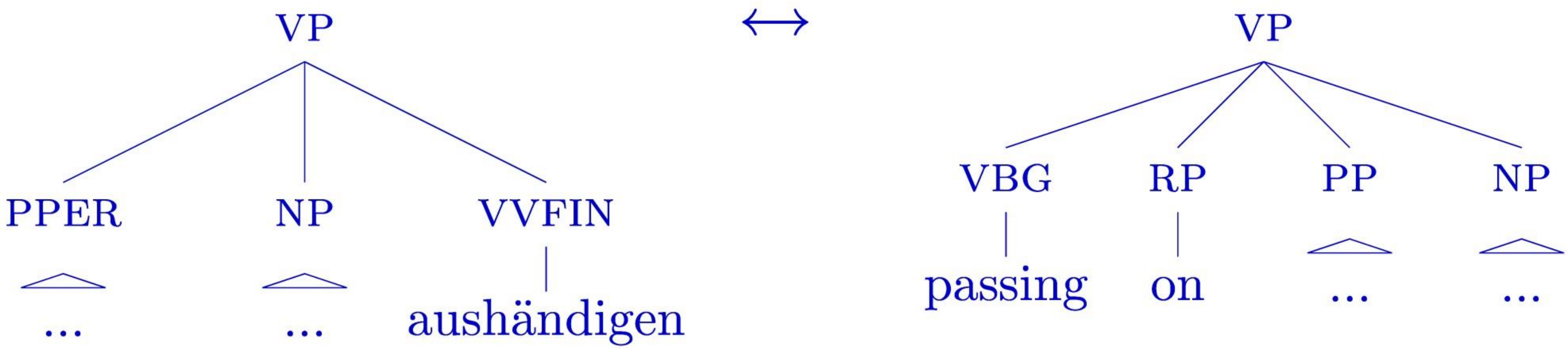


Phrase structure grammar trees with word alignment
(German–English sentence pair.)

Slide from Koehn 2009

Reordering Rule

- Subtree alignment



- Synchronous grammar rule

$$\text{VP} \rightarrow \text{PPER}_1 \ \text{NP}_2 \ \text{aushändigen} \ \mid \ \text{passing on} \ \text{PP}_1 \ \text{NP}_2$$

- Note:
  - one word aushändigen mapped to two words passing on ok
  - but: fully non-terminal rule not possible
    (one-to-one mapping constraint for nonterminals)

Another Rule

- Subtree alignment

- Synchronous grammar rule (stripping out English internal structure)

$$\text{PRO/PP} \rightarrow \text{Ihnen} \;\mid\; \text{to you}$$

- Rule with internal structure

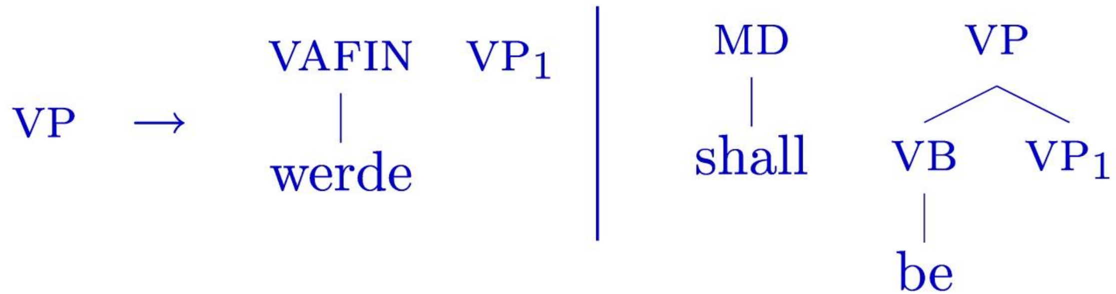
Slide from Koehn 2009

Another Rule

- Translation of German werde to English shall be

- Translation rule needs to include mapping of VP

⇒ Complex rule



Slide from Koehn 2009

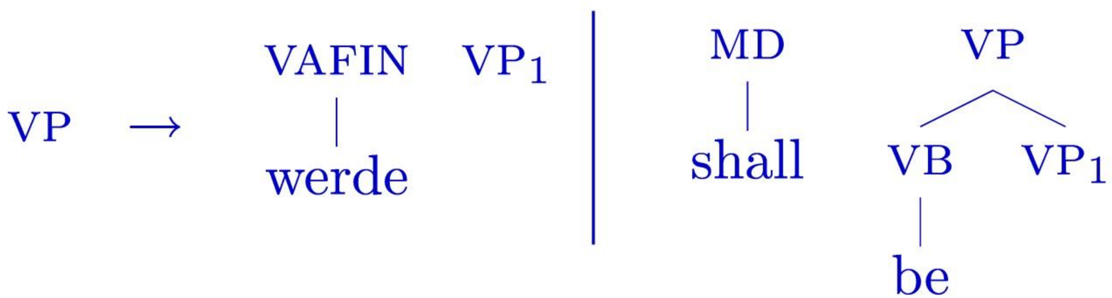
Internal Structure

- Stripping out internal structure

$$\mathrm{VP} \rightarrow \text{werde } \mathrm{VP}_1 \;\mid\; \text{shall be } \mathrm{VP}_1$$

$\Rightarrow$ synchronous context free grammar

- Maintaining internal structure



$\Rightarrow$ synchronous tree substitution grammar

But unfortunately we have some problems

- Two main problems with this approach
  - A text and its translation are not always isomorphic!
  - CFGs make strong independence assumptions

- A text and its translation are not always isomorphic!
  - Heidi Fox looked at two languages that are very similar, French and English, in a 2002 paper
    - Isomorphic means that a constituent was translated as something that can not be viewed as one or more complete constituents in the target parse tree
    - She found widespread non-isomorphic translations
  - Experiments (such as the one in Koehn, Och, Marcu 2003) showed that limiting phrase-based SMT to constituents in a CFG derivation hurts performance substantially
    - This was done by removing phrase blocks that are not complete constituents in a parse tree
    - However, more recent experiments call this result into question

- CFGs make strong independence assumptions
  - With a CFG, after applying a production like S -> NP VP then NP and VP are dealt with independently
  - Unfortunately, in translation with a SCFG, we need to score the language model on the words not only in the NP and the VP, but also across their boundaries
    - To score a trigram language model we need to track two words OUTSIDE of our constituents
    - For parsing (= decoding), we switch from divide and conquer (low order polynomial) for an NP over a certain span to creating a new NP for each set of boundary words!
      - Causes an explosion of NP and VP productions
      - For example, in chart parsing, there will be many NP productions of interest for each chart cell (the difference between them will be the two proceeding words in the translation)

- David Chiang’s Hiero model partially overcomes both of these problems
  - One of very many syntactic SMT models that were published between about 2003 and 2015
  - Work goes back to mid-90s, when Dekai Wu first proposed the basic idea of using SCFGs (not long after the IBM models were proposed)

Chiang: Hierarchical Phrase-based Model

- **Chiang** [ACL, 2005] (best paper award!)
  - context free bi-grammar
  - *one non-terminal* symbol
  - right hand side of rule may include non-terminals and terminals
- *Competitive* with phrase-based models in 2005 DARPA/NIST evaluation

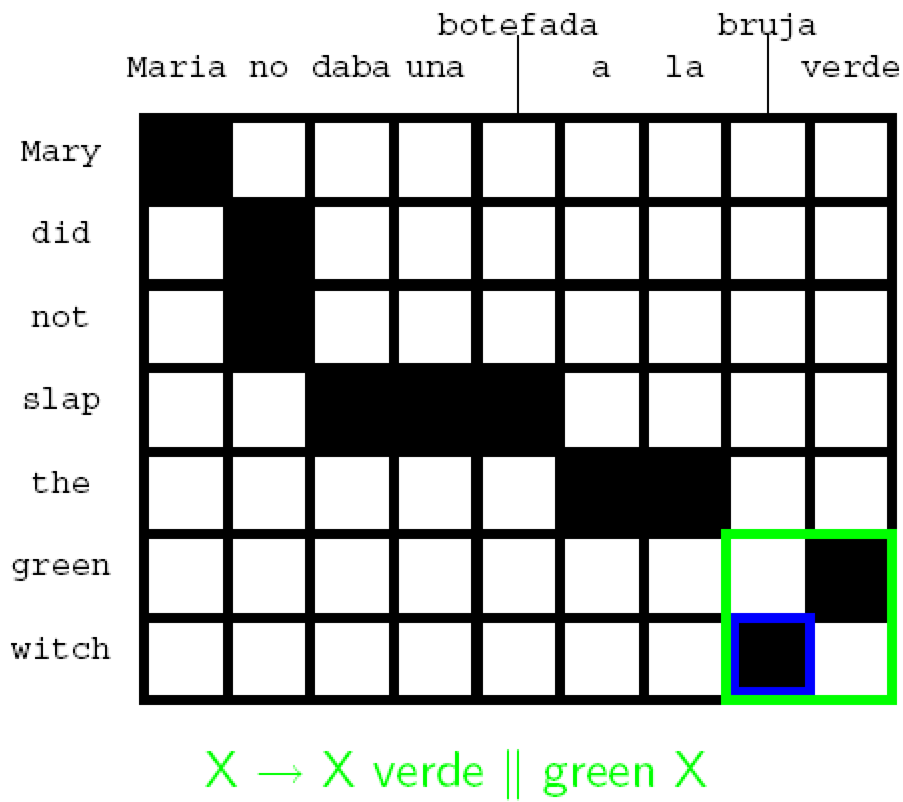
Slide from Koehn and Lopez 2008

Types of Rules

- *Word* translation
  - $X \rightarrow$ *maison* $\|$ *house*
- *Phrasal* translation
  - $X \rightarrow$ *daba una bofetada* $|$ *slap*
- Mixed non-terminal / terminal – *hierarchial phrases*
  - $X \rightarrow X_1$ *bleue* $\|$ *blue* X_1
  - $X \rightarrow$ *ne* X_1 *pas* $\|$ *not* X_1
  - $X \rightarrow X_1\ X_2 \| X_2$ *of* X_1
- Technical rules
  - $S \rightarrow S_1\ X_2 \| S_1\ X_2$
  - $S \rightarrow X_1 \| X_1$

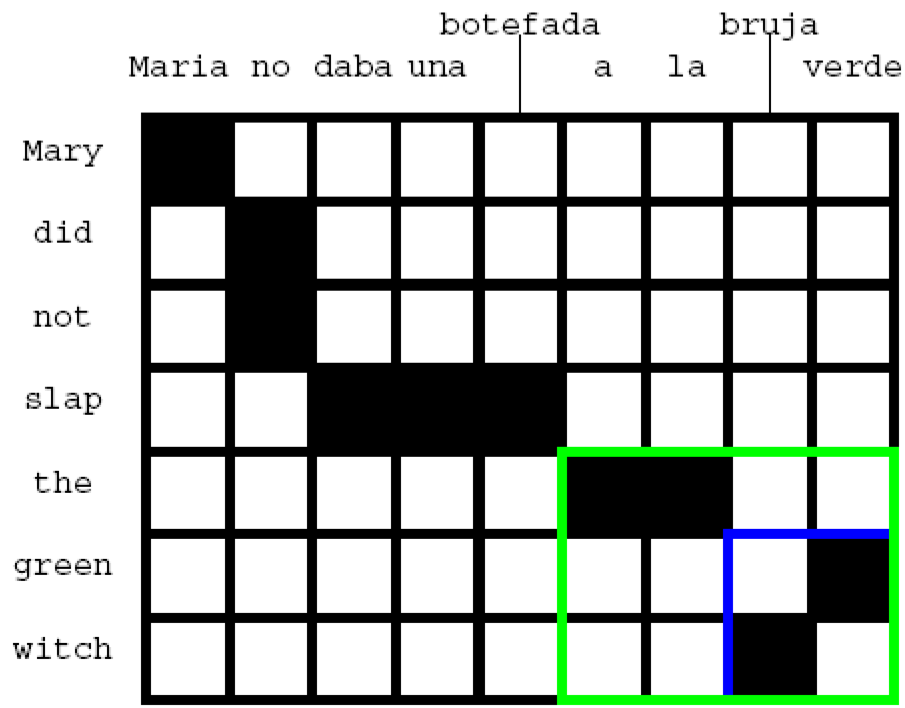
Slide from Koehn and Lopez 2008

Learning Hierarchical Rules



Slide from Koehn and Lopez 2008

Learning Hierarchical Rules



$$X \rightarrow \text{a la } X \parallel \text{the } X$$

Slide from Koehn and Lopez 2008

Comments on Hiero

- Grammar does not depend on labeled trees, and does not depend on preconceived CFG labels (Penn Treebank, etc)
  - Instead, the word alignment alone is used to generate a grammar
  - The grammar contains all phrases that a phrase-based SMT system would use as bottom level productions
  - This does not completely remove the non-isomorphism problem but helps
- Rules are strongly lexicalized so that only a low number of rules apply to a given source span
  - This helps make decoding efficient despite the problem of having to score the language model
- Work in Munich on discriminative models for choosing hierarchical rules has been effective

Comments on Morphology and Syntax in MT

- Phrase-based SMT is robust, and is still state of the art for many language pairs
  - Competitive with or better than rule-based for many tasks (particularly with heuristic linguistic processing)
  - Can be competitive with NMT on some language pairs; but this won't last for much longer
  - Industry workhorse
- Before NMT
  - Many research groups working on taking advantage of syntax in statistical models
  - Hiero is easy to explain, but there are many other models
  - Chinese->English MT (not just SMT) was already dominated by syntactic SMT approaches, a few other language pairs interesting

NMT

- There has been a large amount of work on NMT in the last two years
  - I mostly talked in this lecture about dealing with the poor linguistic assumptions in phrase-based SMT
  - Until NMT appeared, syntactic models thought to be the way forward, now at end?
  - My research group has been working on dealing with morphological richness (particularly in the target language), domain adaptation (out of scope here)
- NMT has changed this in a substantial way
  - For instance, there are a few papers showing that word order doesn't seem to be a major problem in NMT, hurts motivation for syntax
  - Morphological richness is still a problem, but may not need much specialized knowledge in NMT (not known yet)
- 4 areas of work here in Munich
  - Looking at morphological richness and NMT
  - Considering translation problems that were out of reach with SMT (for instance, modelling beyond the sentence level!)
  - Examining character-level models (may help with morphological generalization)
  - Exploiting comparable corpora, particularly for domain adaptation (out of scope here)

- Thanks for your attention!