Statistical Machine Translation
Part V - Advanced Topics

Alex Fraser
Institute for Natural Language Processing
University of Stuttgart

2011.12.09 Seminar: Statistical MT

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

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?

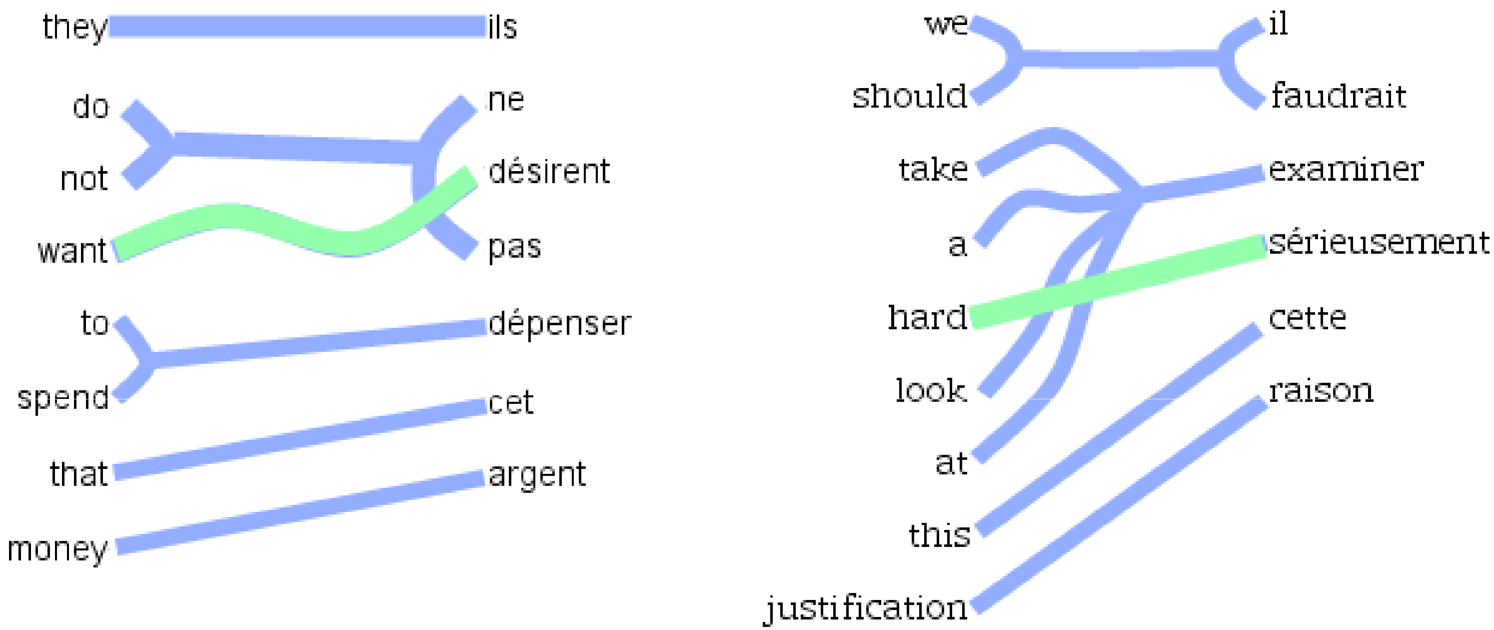
Outline

- 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

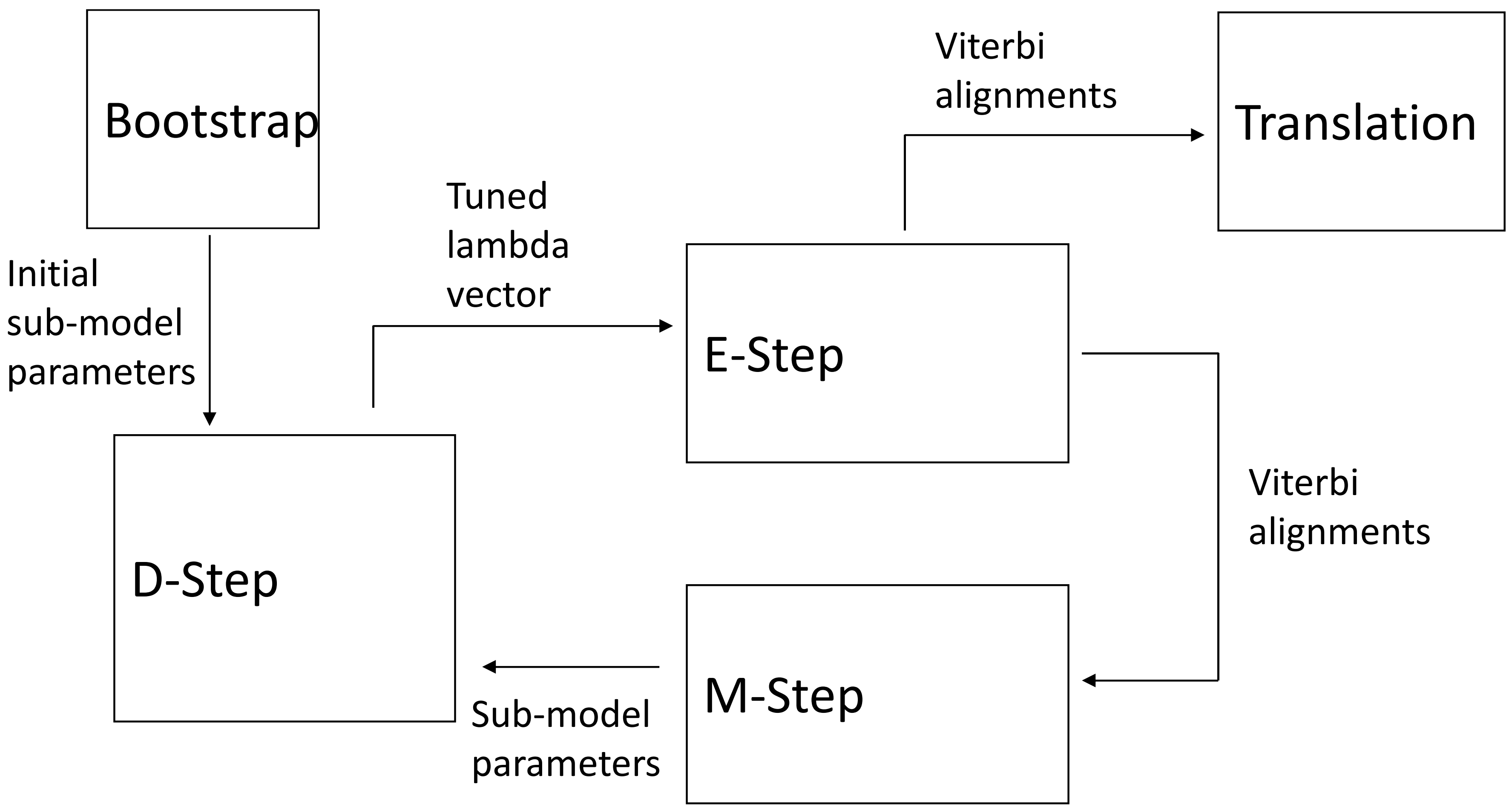
Decomposing LEAF

- 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

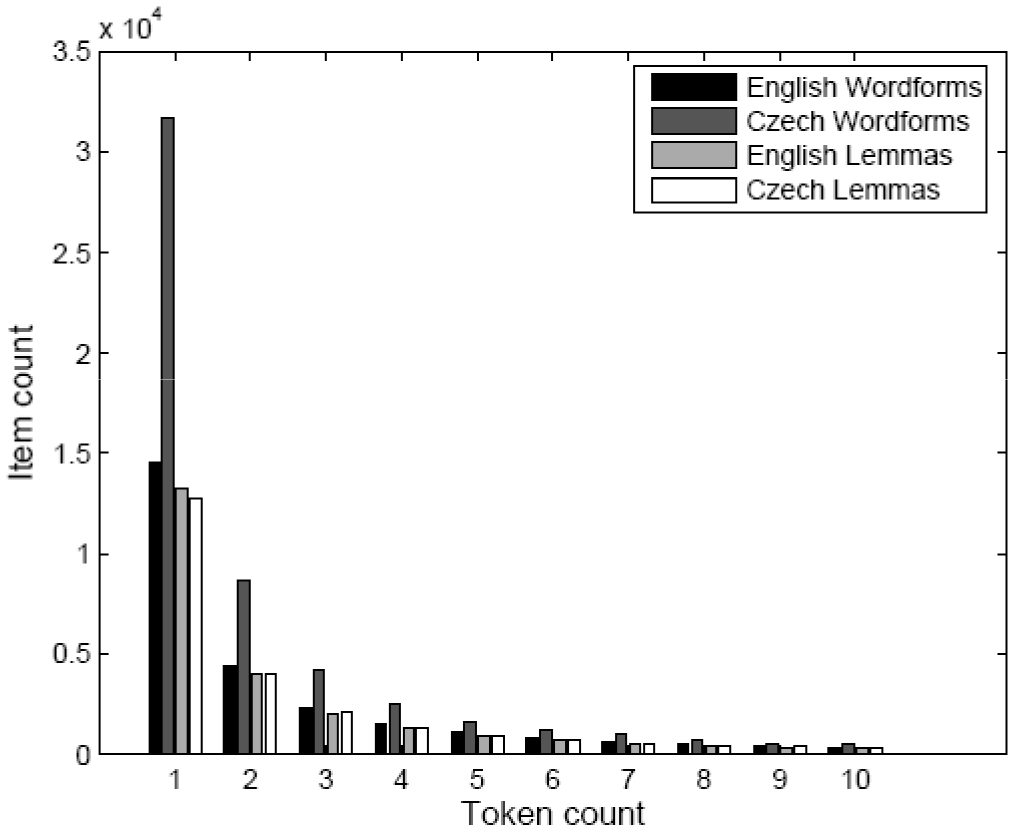
Outlook

- Provides a framework to integrate more morphological and syntactic features in word alignment
  - We are working on this at Stuttgart
  - Other groups doing interesting work using other alignment frameworks (for instance, IBM and ISI for Arabic, Berkeley and ISI for Chinese; many more)

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!
- Last time I presented these slides in 2009:
  - This is not currently improved by using hand-crafted morphological knowledge
  - I doubt this will be the case much longer
- Now: Fabienne Cap has shown using SMOR (Stuttgart Morphological Analyzer) together with corpus statistics is better (Fritzinger and Fraser WMT 2010)

Syntax

- Better modeling of syntax is currently the hottest 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 → English

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

⇒ *Write rules by hand*
  - apply rules to test and training data
  - train standard *phrase-based* SMT system

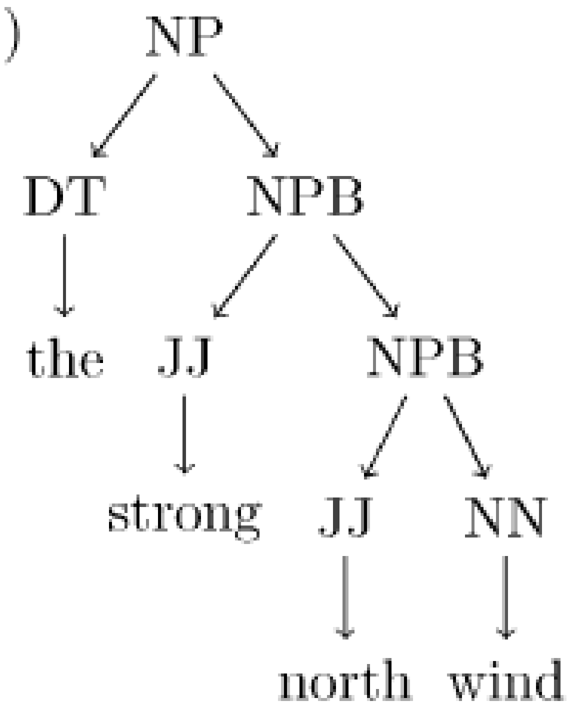
Slide from Koehn and Lopez 2008

But what if we want to integrate probabilities?

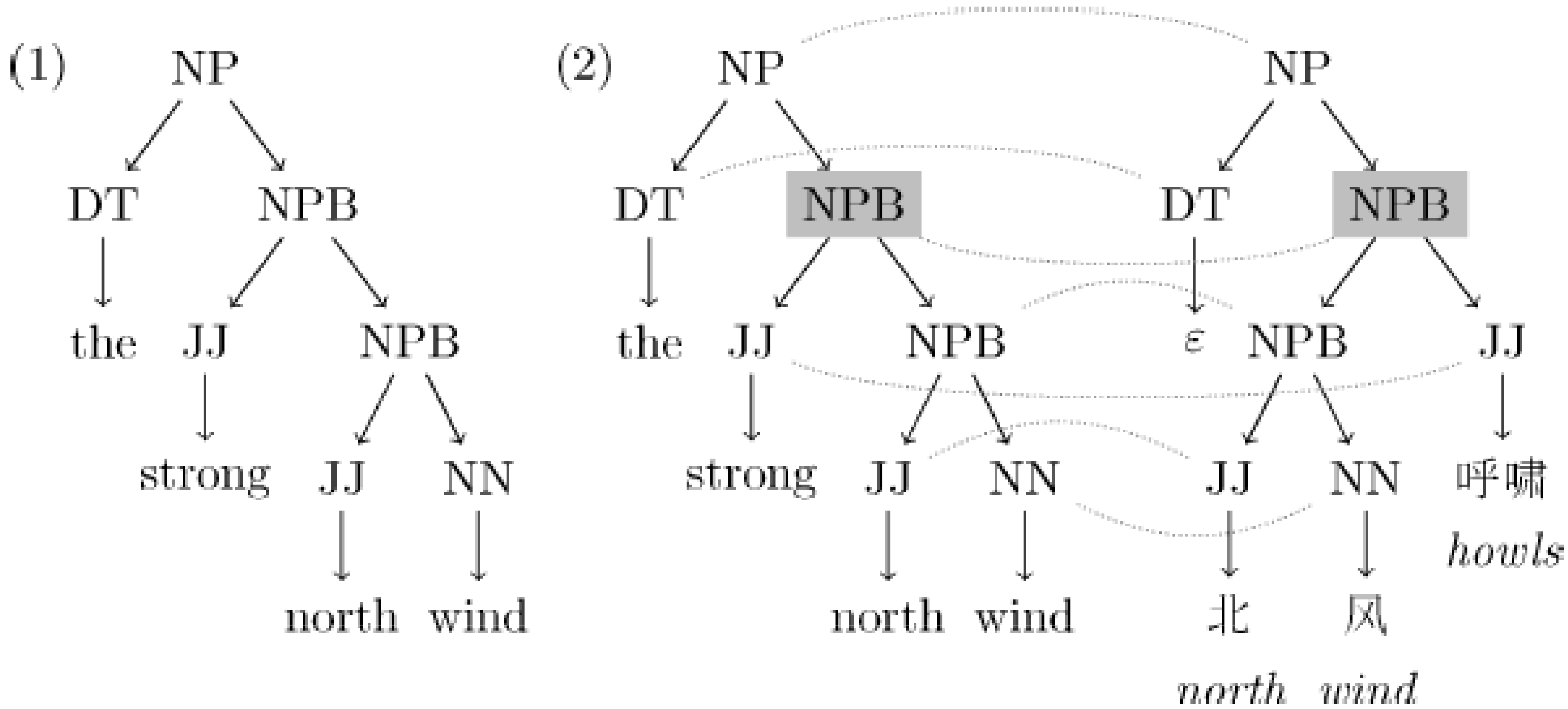
- It turns out that we can!
- We will use something called a synchronous context free grammar (SCFG)
- This is surprisingly simple
  – Just involves defining a CFG with some markup showing what do to with the target language
  – We'll do a short example translating an English NP to a Chinese NP

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

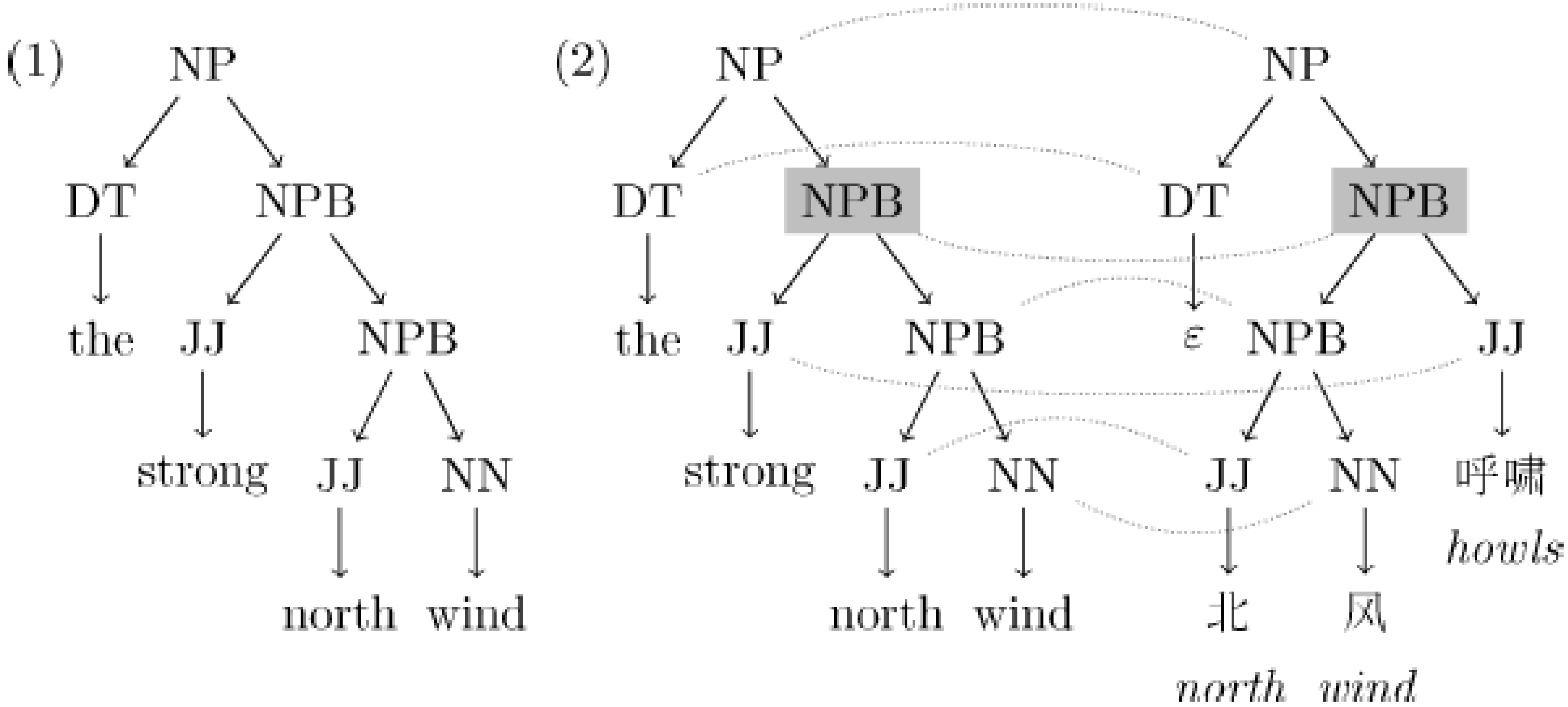
(1)



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

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 have been recently published
  - 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

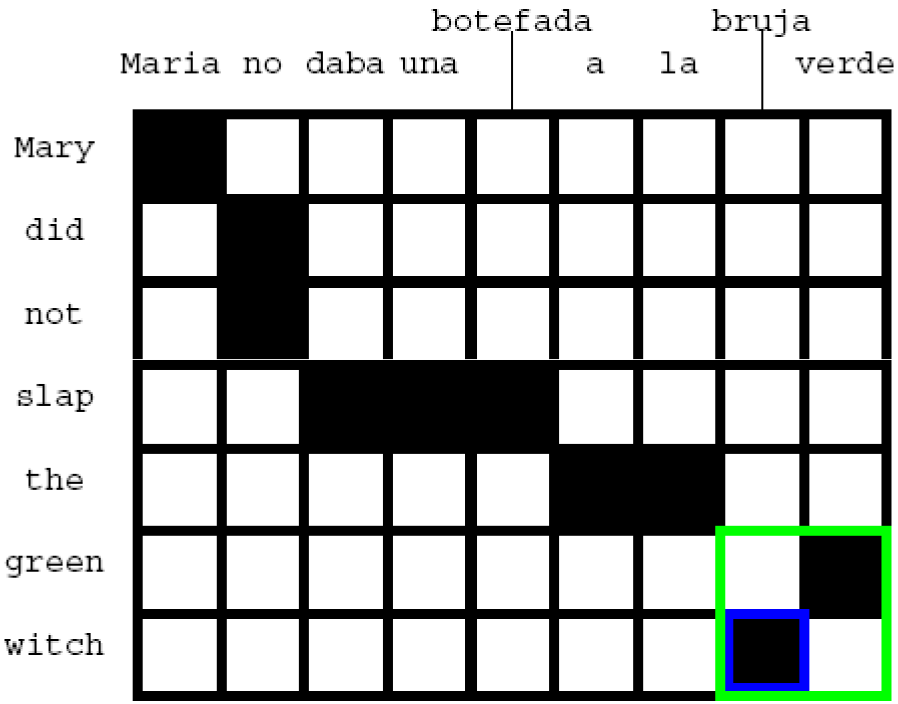
Slide from Koehn and Lopez 2008

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

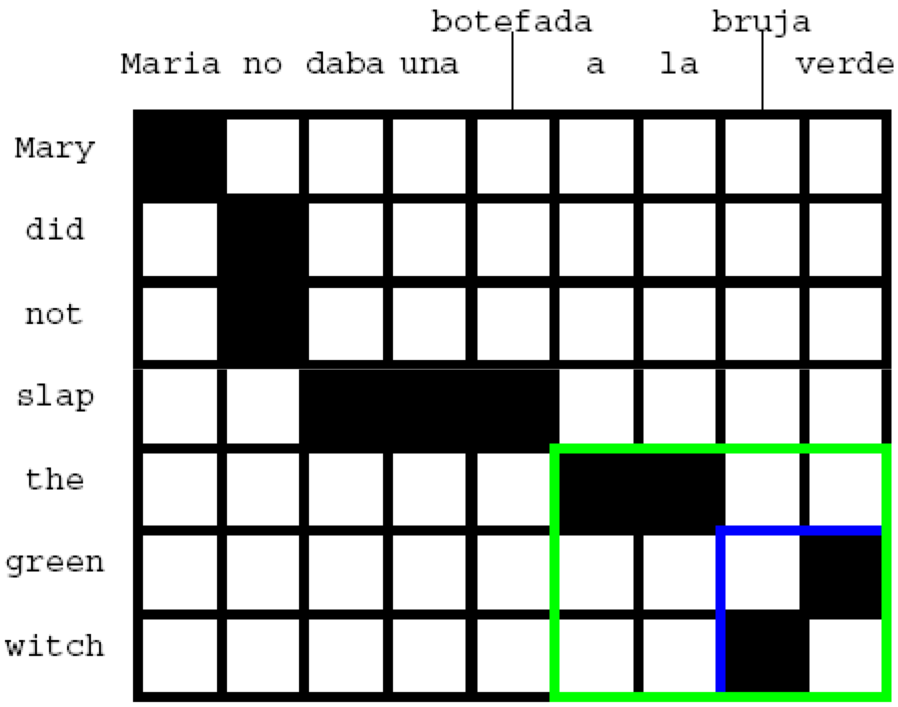
Learning Hierarchical Rules



$X \rightarrow X$ verde || green X

Slide from Koehn and Lopez 2008

Learning Hierarchical Rules



X → a la X || 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

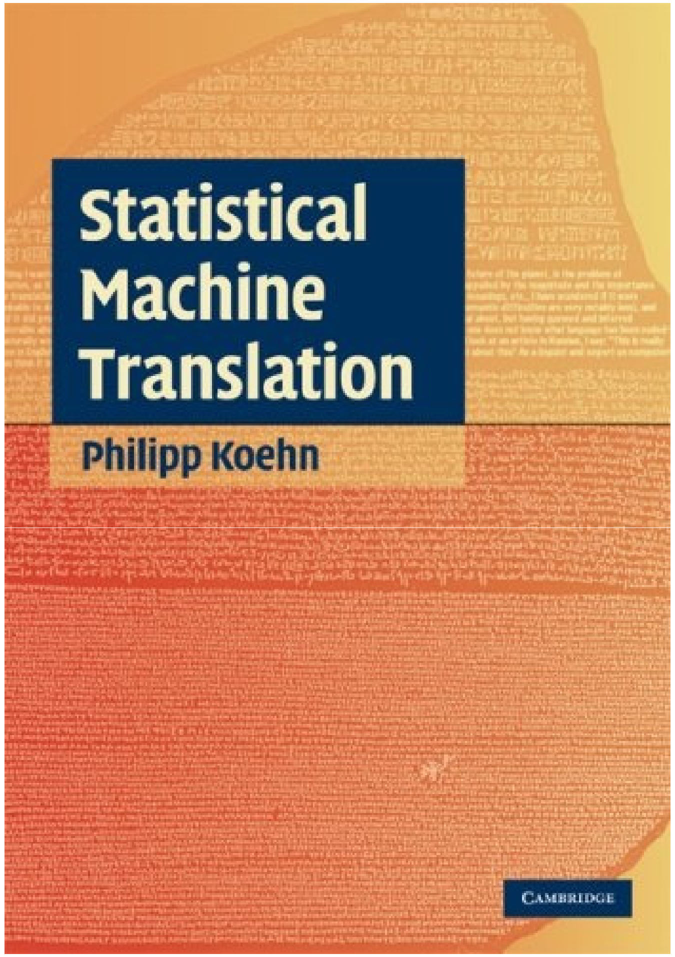
Comments on Morphology and Syntax

- 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)
- Integration of morphological and syntactic models will be the main focus of the next years
  - Many research groups working on this (particularly syntax)
  - Hiero is easy to explain, but there are many others
  - Chinese->English MT (not just SMT) is already dominated by syntactic SMT approaches

Bibliography

- Please see web page for updated version!
- Measuring translation quality
  - Papineni et al 2001: defines BLEU metric
  - Callison-Burch et al 2007: compares automatic metrics including METEOR
- Measuring alignment quality
  - Fraser and Marcu 2007: F-alpha
- Generative alignment models
  - Kevin Knight 1999: tutorial on basics, Model 1 and Model 3
  - Brown et al 1993: IBM Models
  - Vogel et al 1996: HMM model (best model that can be trained using exact EM. See also several recent papers citing this paper)
- Discriminative word alignment models
  - Fraser and Marcu 2007: hybrid generative/discriminative model
  - Moore et al 2006: pure discriminative model

- Phrase-based modeling
  - Och and Ney 2004: Alignment Templates (first phrase-based model)
  - Koehn, Och, Marcu 2003: Phrase-based SMT
- Phrase-based decoding
  - Koehn: manual of Pharaoh (precursor of Moses)
- Syntactic modeling
  - Chiang 2007: unlabeled tree-to-string translation
  - Galley et al 2004: string-to-labeled tree translation
  - Quirk et al 2005: labeled tree-to-string translation
  - Chiang et al 2008, 2009: using labels as soft constraints
  - Many more!
- Surveys of SMT
  - Philipp Koehn 2009: basic textbook (see next slide)
  - Adam Lopez 2008: technical survey of cutting edge


Statistical Machine Translation
Philipp Koehn
CAMBRIDGE

Conclusion

- Lecture 1 covered background, parallel corpora, sentence alignment, evaluation and introduced modeling
- Lecture 2 was on word alignment using both exact and approximate EM
- Lecture 3 was on phrase-based modeling and decoding
- Lecture 4 was on log-linear models and MERT
- Lecture 5 briefly touched on new research areas in word alignment, morphology and syntax

- Thanks for your attention!