Machine Learning Approaches for Natural Language Processing

WANG Haifeng （王海峰）

Toshiba (China) R&D Center

MLA, Nov. 7, 2009

Outline

- Overview of Natural Language Processing (NLP)
- Machine Learning Approaches for NLP
- Overview of Machine Translation (MT)
- Semi-Supervised Boosting for Statistical Word Alignment and SMT

CL vs. NLP

Computational Linguistics, CL	Natural Language Processing, NLP
ACL: Association for Computational Linguistics	EMNLP: Empirical Methods in Natural Language Processing
COLING: International Conference on Computational Linguistics	IJCNLP: International Joint Conference on Natural Language Processing
ICCL： International Committee on Computational Linguistics	AFNLP: Asian Federation of Natural Language Processing
CNCCL： Chinese National Conference on Computational Linguistics	YSSNLP: Young Scholar Symposium on Natural Language Processing
ICL： Institute of Computational Linguistics	**NLPLAB: **Natural Language Processing LAB

Impact?

History

Theory

Methodology

NLP Areas

- ACL-IJCNLP 2009

Area	#Submission	#Accepted	Rate
Machine Translation	82	23	28.0%
Semantics	67	14	20.9%
Syntax and Parsing	49	14	28.6%
Information Extraction	49	10	20.4%
Discourse, Dialogue and Pragmatics	43	9	20.9%
Summarization and Generation	44	8	18.2%
Phonology, Morphology, Segmentation, POS, Chunking	31	8	25.8%
Sentiment Analysis, Opinion Mining, Classification	45	7	15.6%
Statistical and Machine Learning Methods	40	6	15.0%
Spoken Language Processing	19	6	31.6%
Information Retrieval	28	4	14.3%
Language Resource	26	4	15.4%
Text Mining and NLP Applications	21	4	19.0%
Question Answering	25	3	12.0%
Total	569	120	21.1%

NLP Areas

Application

- ACL-IJCNLP 2009

Area	#Submission	#Accepted	Rate
Machine Translation	82	23	28.0%
Semantics	67	14	20.9%
Syntax and Parsing	49	14	28.6%
Information Extraction	49	10	20.4%
Discourse, Dialogue and Pragmatics	43	9	20.9%
Summarization and Generation	44	8	18.2%
Phonology, Morphology, Segmentation, POS, Chunking	31	8	25.8%
Sentiment Analysis, Opinion Mining, Classification	45	7	15.6%
Statistical and Machine Learning Methods	40	6	15.0%
Spoken Language Processing	19	6	31.6%
Information Retrieval	28	4	14.3%
Language Resource	26	4	15.4%
Text Mining and NLP Applications	21	4	19.0%
Question Answering	25	3	12.0%
Total	569	120	21.1%

NLP Taxonomy

- Sub-task
  - Analysis & understanding, generation
- Level
  - Morphology, syntax, semantics, pragmatics
- Grammar
  - PS, DS, LFG, HPSG, CCG …
- Unit
  - Character, word, phrase, sentence, paragraph …
- Style
  - Spoken language, written language
- Application
  - Translation, information retrieval and extraction, sentiment, QA, summarization, grammar check …
- Approach
  - Rationalist and empiricist approaches
- Data
  - Lexicon, rules, corpus (labeled and unlabeled)

Difficulties

- Complex structure
  - Mapping between string and structure
- Ambiguities
  - Disambiguation
- Examples
  - 打：打酱油、打毛衣、打人、打针 ……
  - pretty little girls' school
    - Does the school look little?
    - Do the girls look little?
    - Do the girls look pretty?
    - Does the school look pretty?

Approaches

- Rationalist approaches
  - Linguistic theory
  - Grammar system
  - Rules
    - Usually manually compiled
  - Popular in NLP application (e.g. RBMT)

It must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

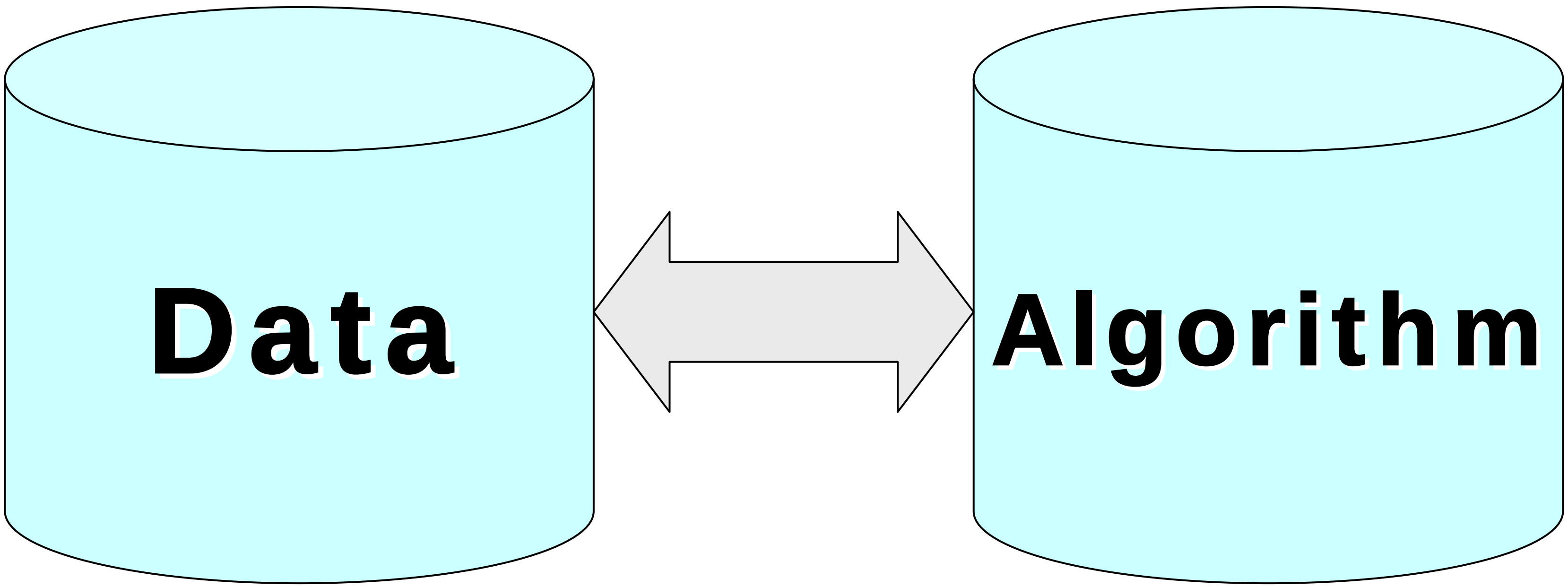
Noam Chomsky

- Empiricist approaches
  - Corpus
    - Labeled, unlabeled
    - Monolingual, multilingual
  - Statistical and Machine Learning Approaches
  - Dominant approach in NLP research

Whenever I fire a linguist our system performance improves.

Frederick Jelinek

Data vs. Algorithm



Outline

- Overview of Natural Language Processing (NLP)
- Machine Learning Approaches for NLP
- Overview of Machine Translation (MT)
- Semi-Supervised Boosting for Statistical Word Alignment and SMT

Why ML in NLP

- Practical concerns
  - Data
    - More and more unlabeled language data available on the web and elsewhere
    - Labeled data are easier to create than rules
  - Method
    - It is hard for knowledge engineering methods to cope with the growing flood of data
    - Machine learning can be used to automate knowledge acquisition and inference
  - Computing resource
    - More and more powerful (Moore’s law)
- Theoretical contribution
  - Reasonably solid foundations (theory and algorithms)

ML gives elegant, well-founded solutions to NLP problems

NLP comes with data and gives meaning to ML's math

Designing ML for NLP

- Data
  - How to access and use data
- Target function
  - Concept to be learnt
- Representation
  - Representation of hypotheses
  - Representation of data
- Learning algorithm
  - Conditioned to the representation
  - Inductive learning assumption

ML applications in NLP

- NLP Areas
  - Machine Translation
  - Semantics
  - Syntax and Parsing
  - Information Extraction
  - Information Retrieval
  - Summarization and Generation
  - Question Answering
  - ……
  - ML methods have been used in most NLP areas

- ML Methods
  - HMM, ME, CRF, SVM, Boosting, Co-training ……
  - Many ML methods have been or will be used in NLP

ML at Leading NLP Conferences

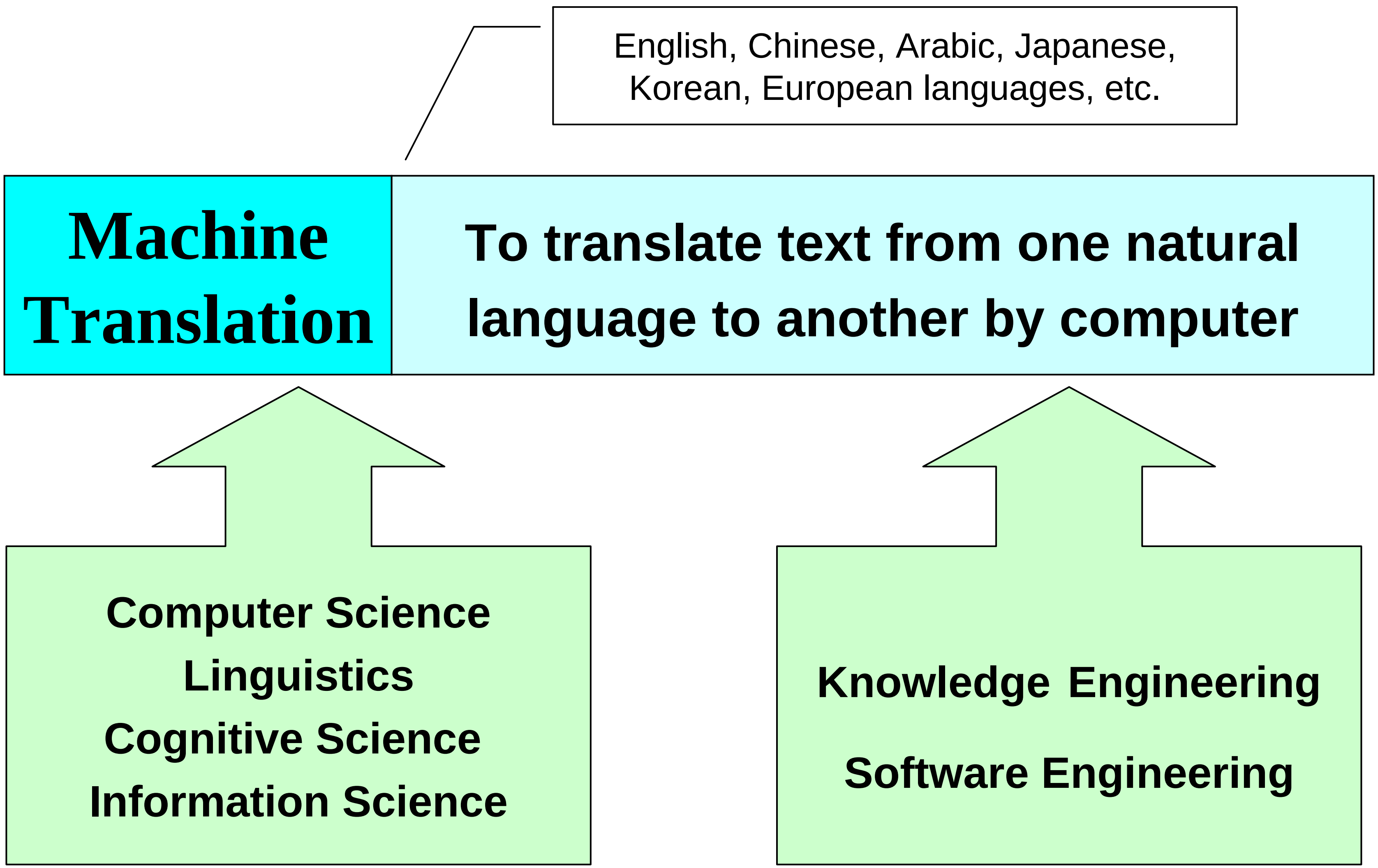
ACL-IJCNLP 2009

- 2 sessions on “Statistical and Machine Learning Methods” (1 best paper)
  - Improving learning method
  - Mapping Instructions to Actions (one of the best papers)
  - Semantic
  - Word Segmentation
  - POS
- Much more ML related papers in other sessions
  - Invited talk
    - Heterogeneous Transfer Learning with Real-world Applications (Qiang Yang)
  - Machine translation
  - Parsing
  - Semantics
  - Question Answering
  - Information Extraction
  - ……

Outline

- Overview of Natural Language Processing (NLP)
- Machine Learning Approaches for NLP
- Overview of Machine Translation (MT)
- Semi-Supervised Boosting for Statistical Word Alignment and SMT

Machine Translation



Applications

Purpose

- Information distribution
- Information gathering
- Communication
- CLIR

Unit

- Word
- Phrase
- Sentence
- Paragraph

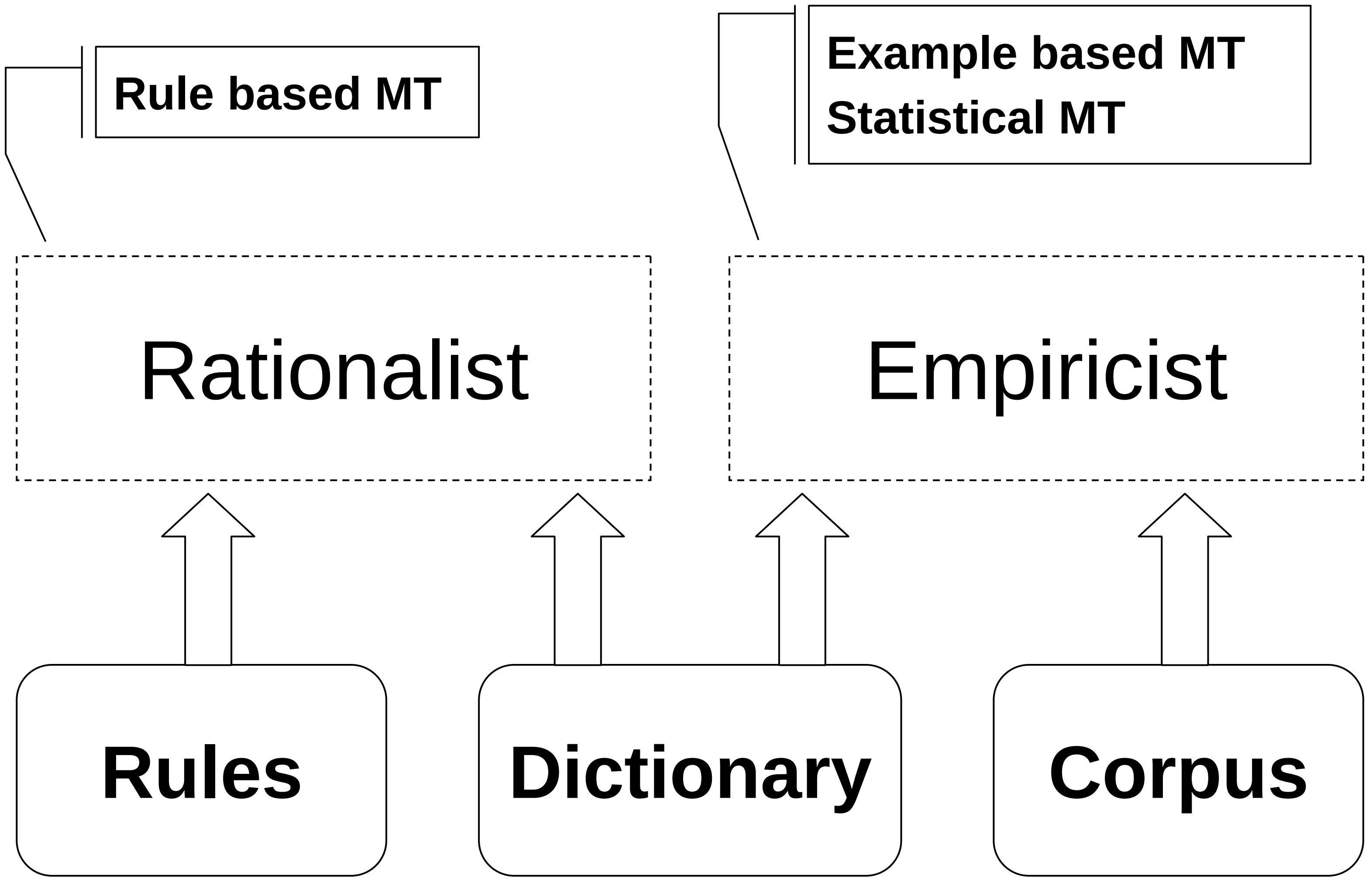
Translation Mode

- Automatic MT
- Computer Aided Translation
- Speech Translation

Product Form

- Software package
- MT engine license
- Translation service
- Hardware preinstall

How


Rule based MT
Example based MT
Statistical MT
Rationalist
Empiricist
Rules
Dictionary
Corpus

Methods

RBMT

- Good at describing linguistic phenomena
- Dominant method in commercial product
- Potential for Improvement (Zhu. 2005)

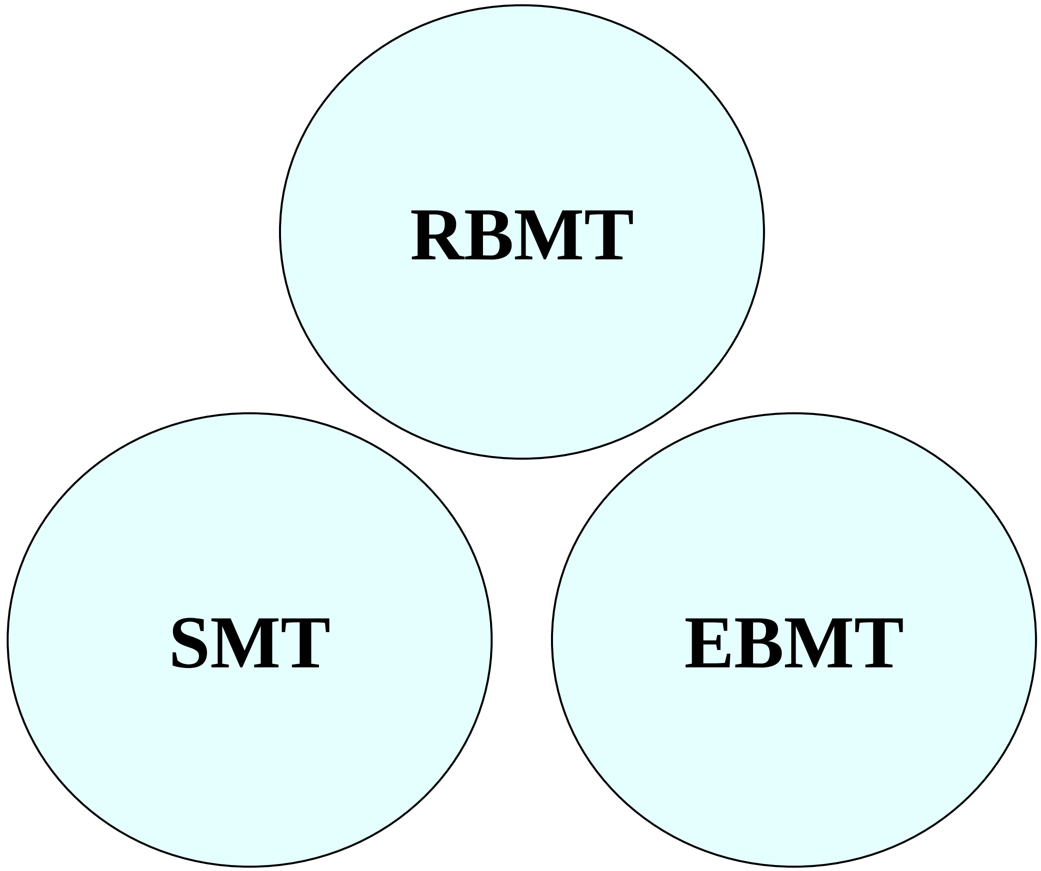
EBMT

- Good at translating similar sentence
- Still popular (MT Journal, MT Summit)
- Comparable with SMT (Way 2005, Liu 2006)

SMT

- Model, learning, robustness
- Dominant in MT research
- Won NIST MT evaluations

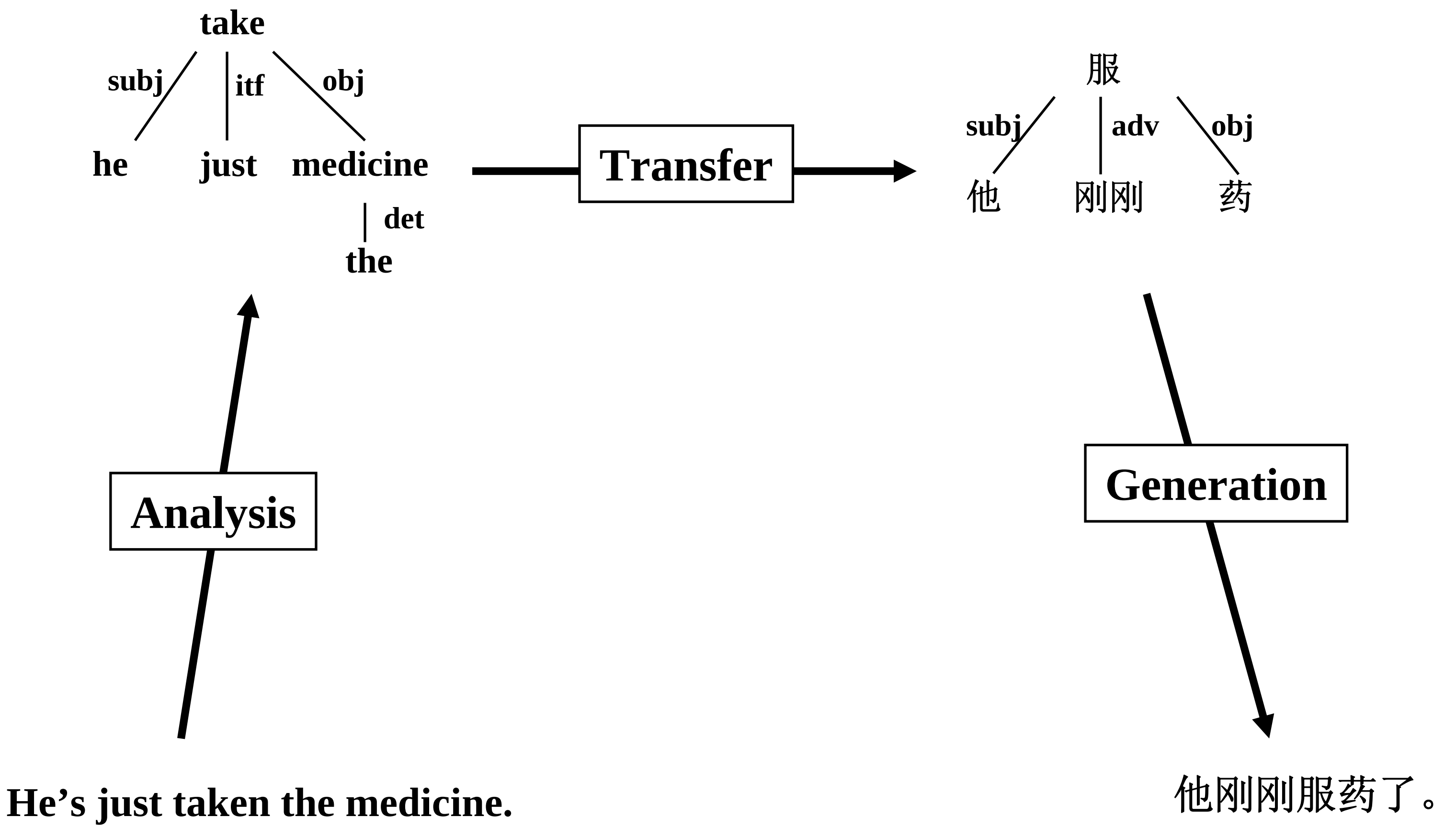
Combination



Wu, Chiang, Groves, Liu

RBMT

Transfer-based RBMT



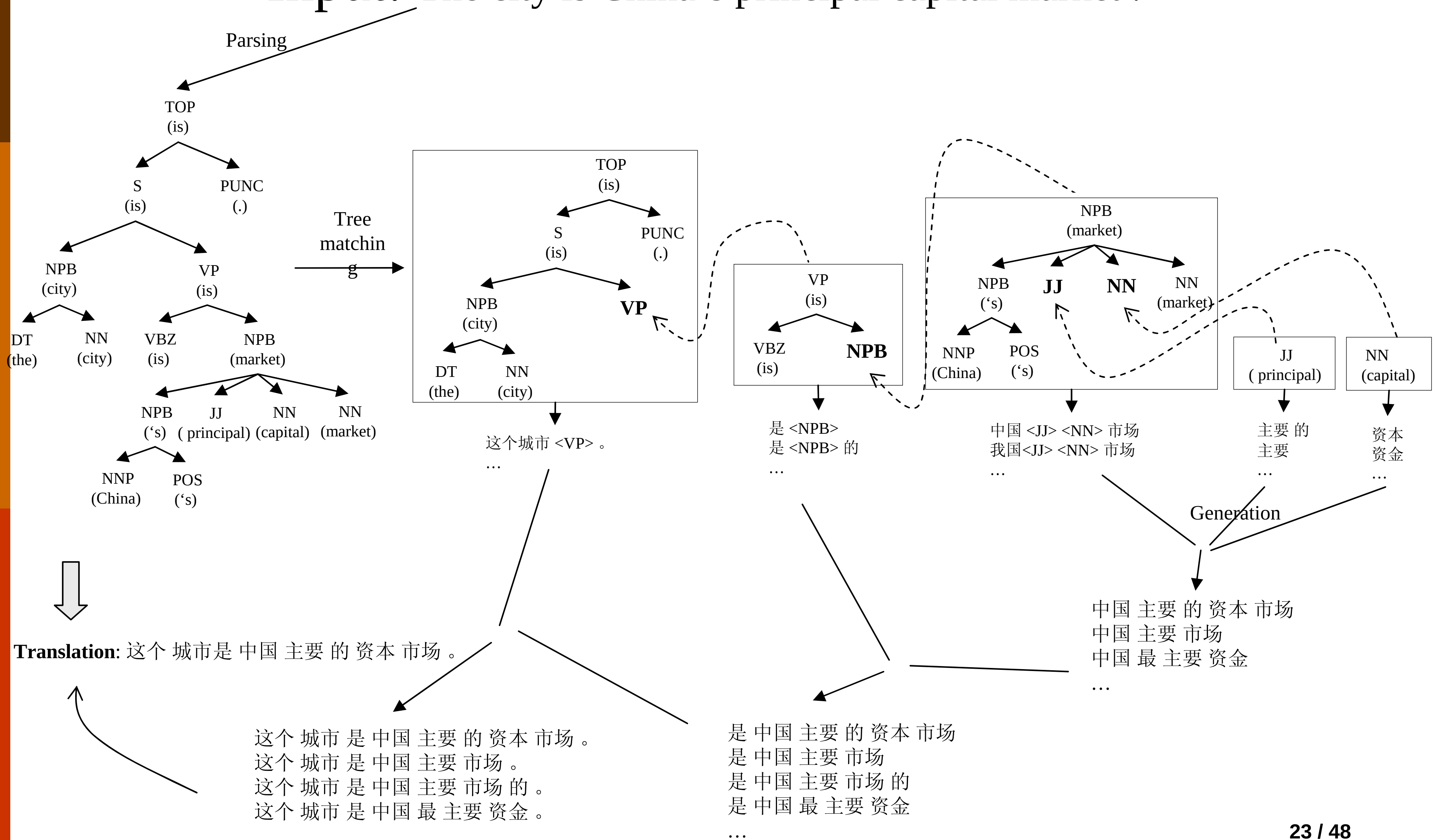
- Hierarchical
- Fine grained
- Scalable

EBMT

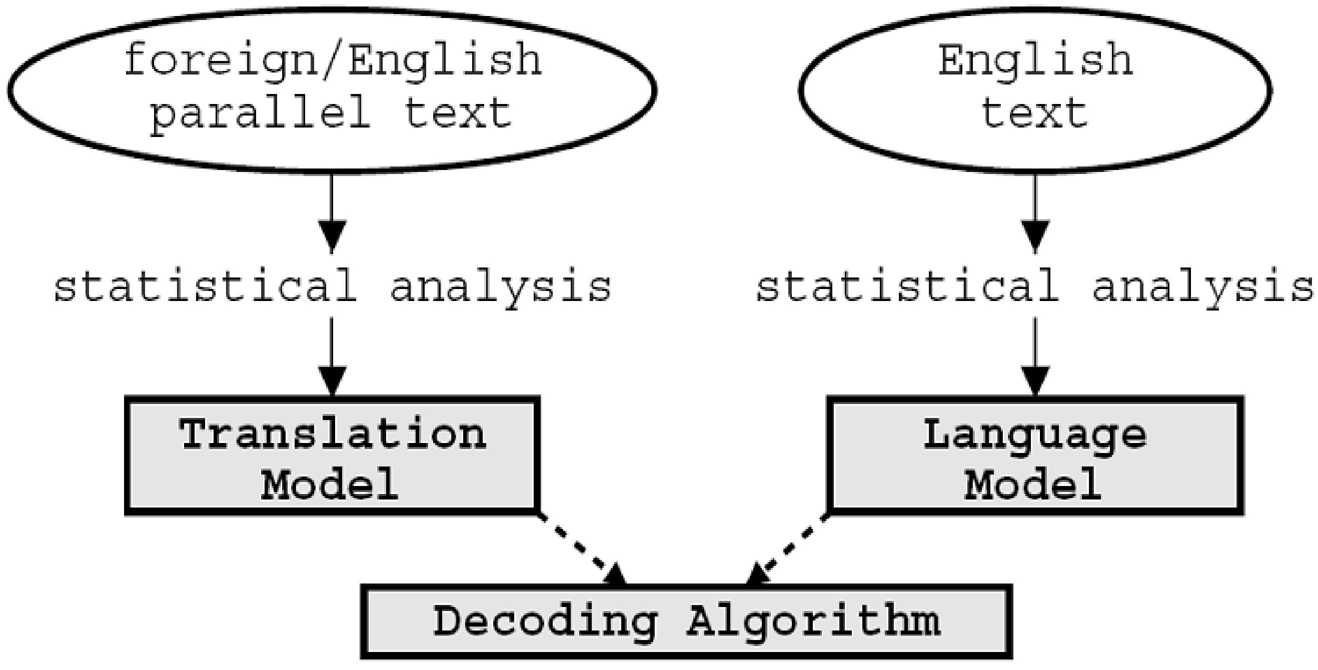
- Example-based machine translation (EBMT)
  - Machine translation by example-guided inference, or machine translation by the analogy principle (Nagao, 1984)
- Three main components
  - Match fragments against a database of real examples
  - Identify the corresponding translation fragments
  - Combine these to give the final translation
- Performance
  - Good performance in domain specific application

Tree Based EBMT

Input: The city is China’s principal capital market .



Overview of SMT



- Word-based SMT (Brown et al., 1990 & 1993)
- Phrase-based SMT (Koehn et al., 2003)
- Syntax-based SMT (Wu, 1997; Chiang, 2005)

Method

- Statistical theory

$$\hat{e}_1^I = \arg\max_{e_1^I} \{ \Pr(e_1^I \mid f_1^J) \}$$

- Generative method
  - Translation process is broken down into steps
  - Each step is modeled by a probability distribution
  - Each probability distribution is estimated by maximum likelihood
- Discriminative method
  - Model consists of a number of features
  - Each feature has a weight
  - Feature weights are optimized on development set

Model

- Source-channel model

$$\hat{e}_1^I = \arg\max_{e_1^I}\{\Pr(e_1^I)\cdot \Pr(f_1^J \mid e_1^I)\}$$

- Log-linear model

$$\hat{e}_1^I = \arg\max_{e_1^I}\left\{\frac{\exp[\sum_{m=1}^{M}\lambda_m h_m(e_1^I, f_1^J)]}{\sum_{e_1'^I}\exp[\sum_{m=1}^{M}\lambda_m h_m(e_1'^I, f_1^J)]}\right\}$$

$$= \arg\max_{e_1^I}\{\exp[\sum_{m=1}^{M}\lambda_m h_m(e_1^I, f_1^J)]\}$$

Word-based SMT – IBM Model 1

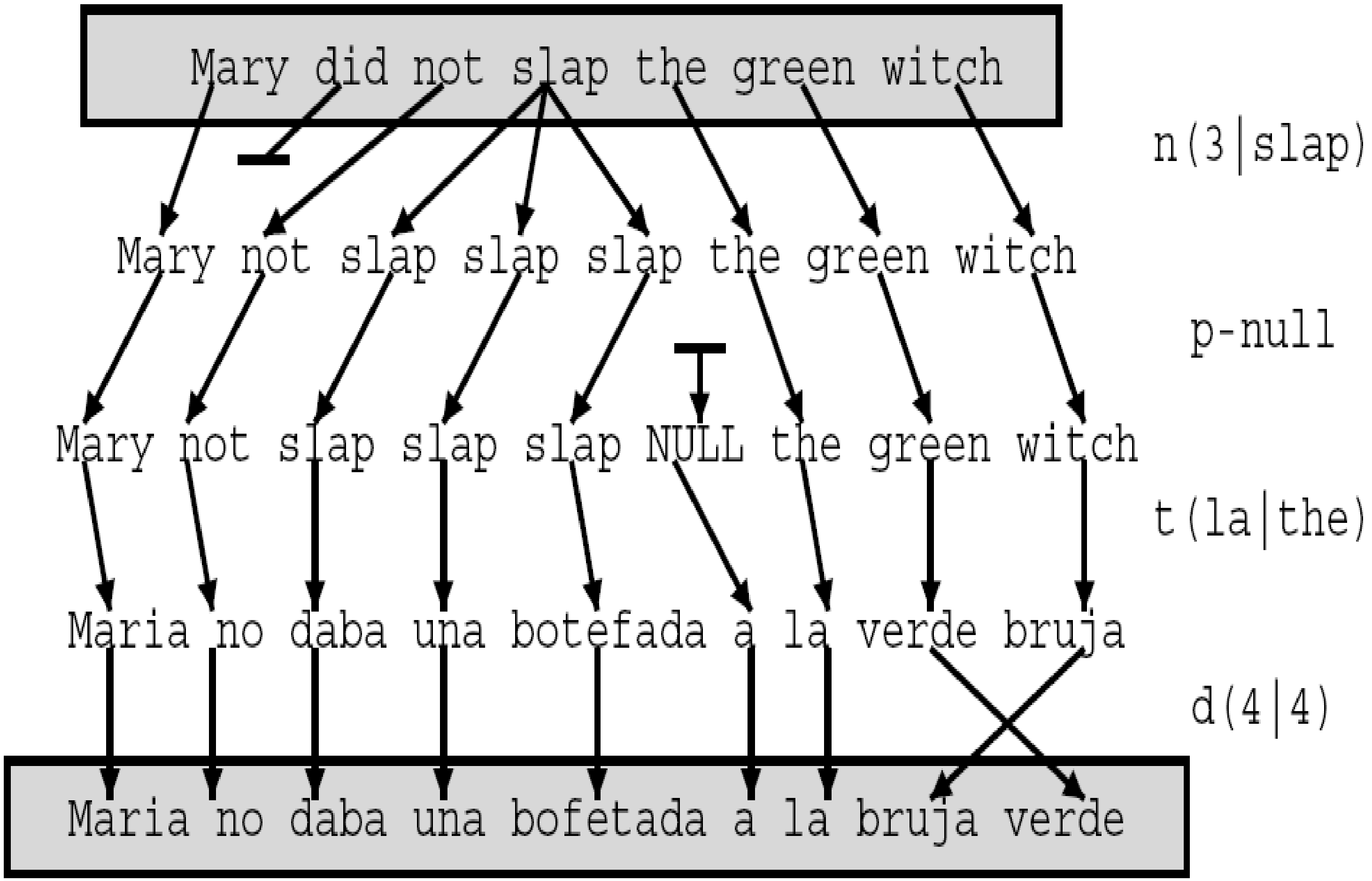
- Only uses lexical translation
- Lexical translation probabilities is estimated from a parallel corpus
- Chicken and egg problem
  - if we had the alignments,
    - we could estimate the parameters of our generative model
  - if we had the parameters,
    - we could estimate the alignments
- EM algorithm
  - Initialize model parameters (e.g. uniform)
  - Assign probabilities to the missing data
  - Estimate model parameters from completed data
  - Iterate

Word-based SMT – 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

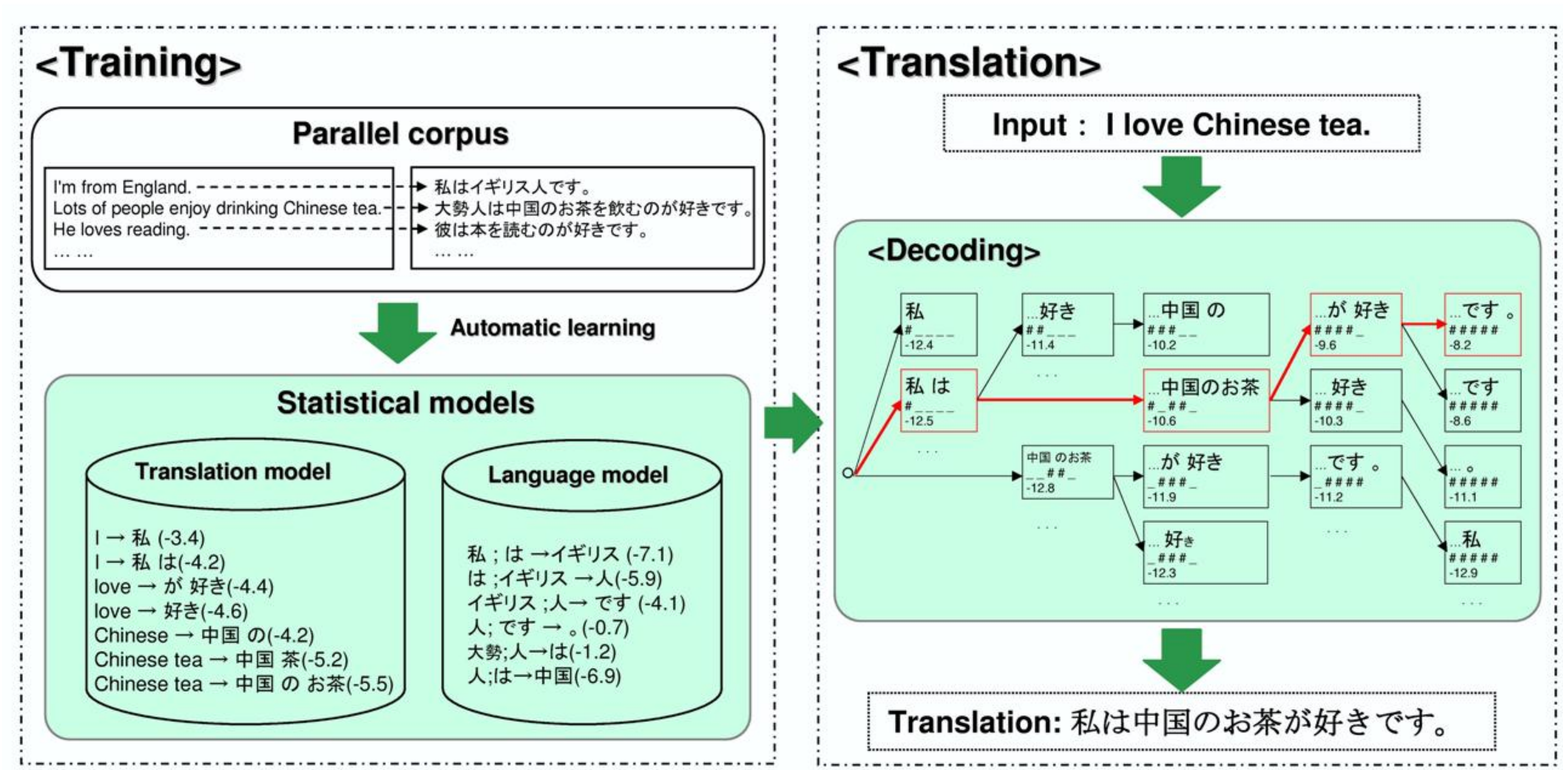
- Training of a higher IBM model builds on previous model

Word-based SMT – IBM Model 3



From (Knight and Koehn)

Phrase-based SMT

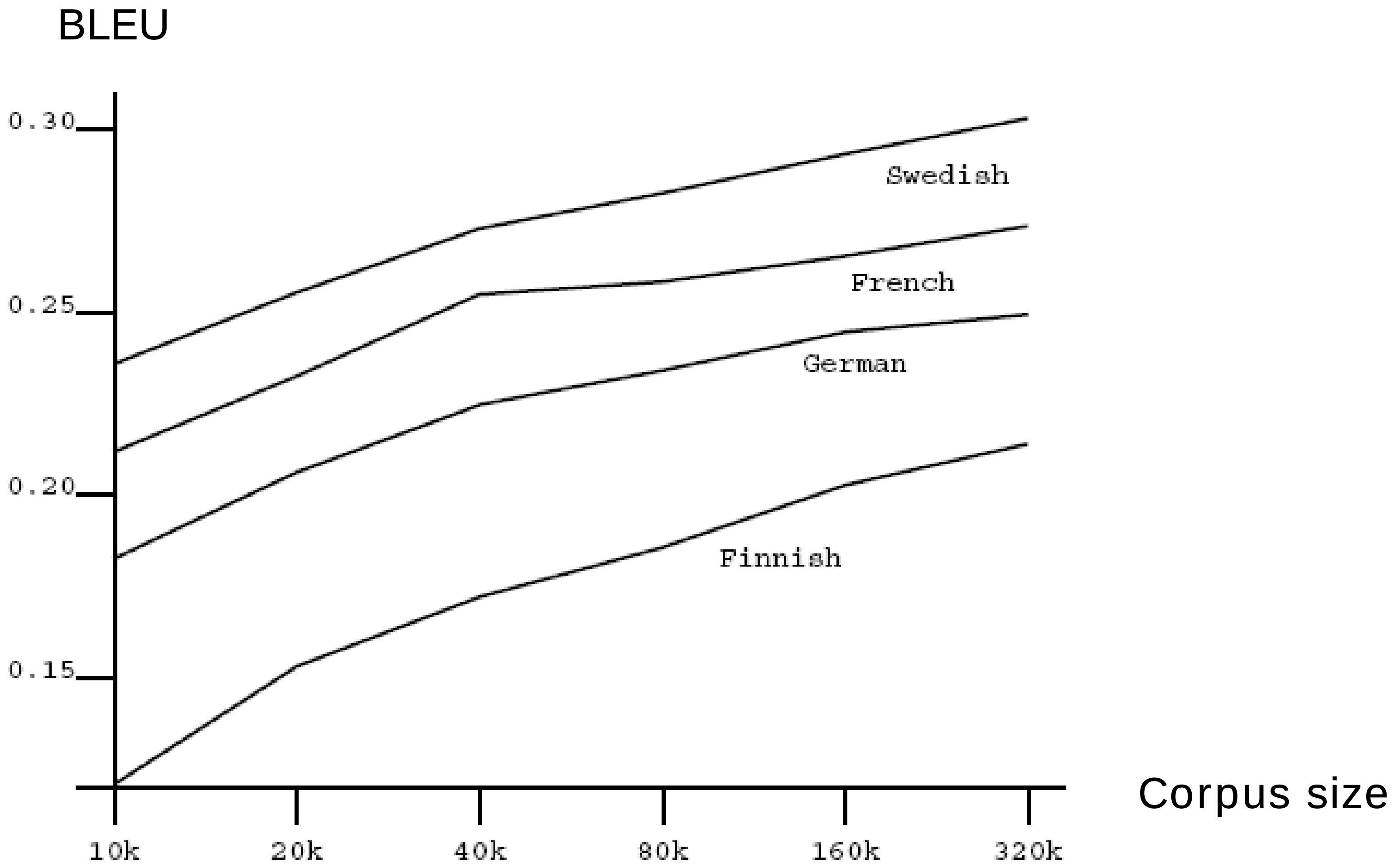


- Foreign input is segmented in phrases
  - Not necessarily linguistically motivated
- Each foreign phrase is translated into native phrase
  - Search a phrase table
- Phrases are reordered

Syntax-based SMT

- Why syntax in SMT
  - More grammatical output
  - Syntax aware re-ordering
  - Accurate insertion of function words
- Grammars
  - Synchronous Context Free Grammars (SCFG)
  - Linguistically informed grammars
- Models
  - Tree-to-String
  - String-to-Tree
  - Tree-to-Tree
- Disadvantage

More Data, Better Translations



From (Koehn, 2003: Europarl Corpus)

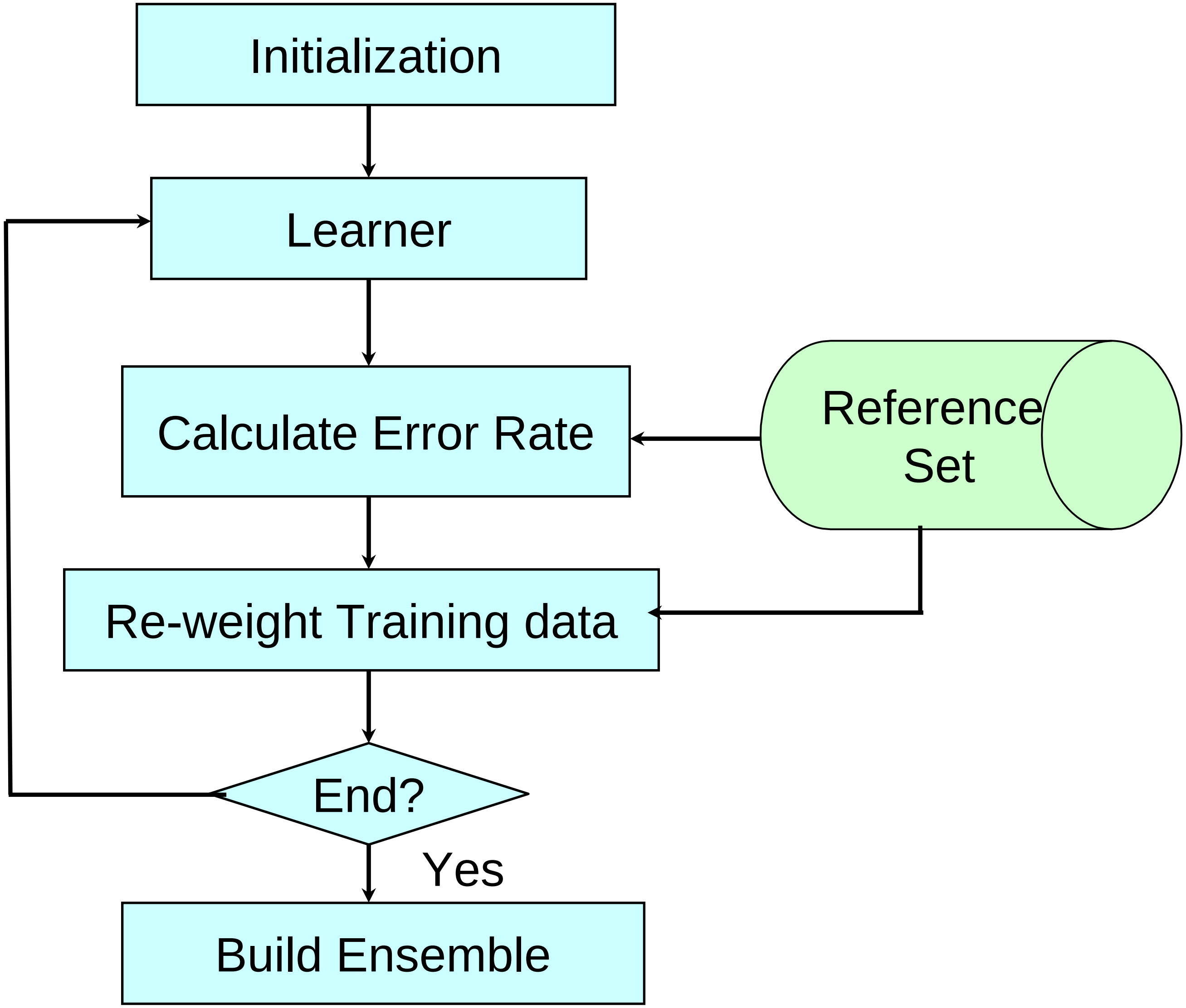
Outline

- Overview of Natural Language Processing (NLP)
- Machine Learning Approaches for NLP
- Overview of Machine Translation (MT)
- Semi-Supervised Boosting for Statistical Word Alignment and SMT

Motivation

- Using both labeled and unlabeled data
- Unlabeled data
  - Aligned automatically
  - IBM models
  - Large, easy
- Labeled data
  - Aligned by human
  - Following a given alignment standard
  - Small, hard
- Adjust automatic alignment results by using manually aligned data

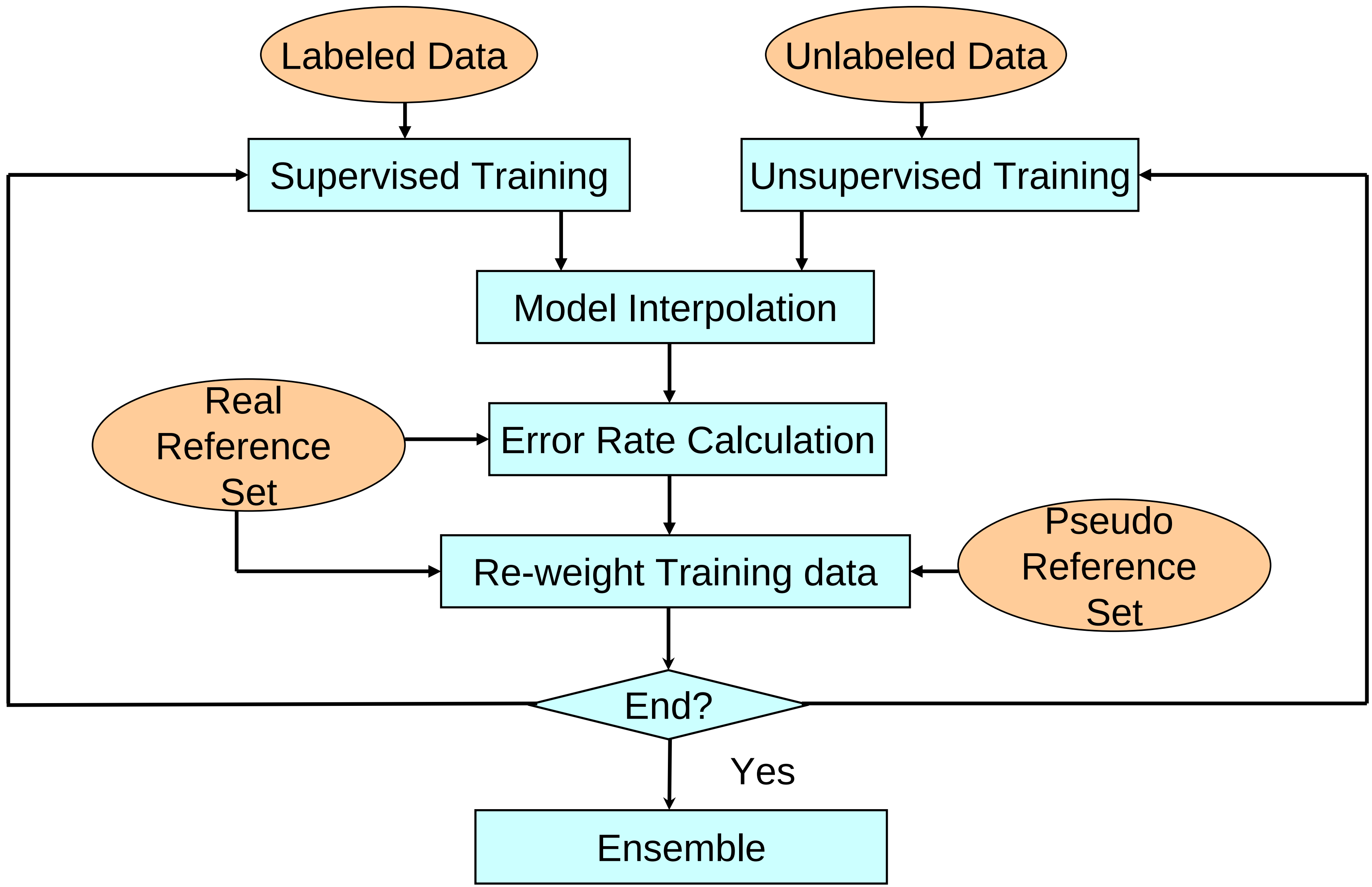
Boosting


Initialization
Learner
Calculate Error Rate
Reference
Set
Re-weight Training data
End?
Yes
Build Ensemble

Semi-Supervised Boosting

- Three main problems
  - Semi-supervised learner
    - Combine labeled data and unlabeled data
  - Reference set
    - Automatically construct a reference set for unlabeled data
  - Error rate calculation
    - How to calculate the error rate with both labeled data and unlabeled data

Semi-Supervised Boosting Word Alignment



Word Alignment Model

- Supervised alignment model
  - Calculate the probabilities for IBM Model based on the labeled data

- Unsupervised alignment model
  - EM training for IBM Model

- Perform model interpolation

$$\Pr(\mathbf{a}, \mathbf{f} \mid \mathbf{e})$$
$$= \lambda \cdot \Pr_S(\mathbf{a}, \mathbf{f} \mid \mathbf{e}) + (1 - \lambda) \cdot \Pr_U(\mathbf{a}, \mathbf{f} \mid \mathbf{e})$$

Pseudo Reference Set Construction

- Obtain bi-directional word alignment sets S_1 and S_2 on the training data
- Obtain the intersection set of these two alignment sets $$I = S_1 \cap S_2$$
- Filter the union set of the two alignment sets

$$C = \{(s,t) \mid p(t \mid s) > \delta_1 \,\&\, count(s,t) > \delta_2\}$$

where $$p(t \mid s) = \frac{count(s,t)}{\sum_t count(s,t)}$$

- Build the pseudo reference set

$$R = I \cup C$$

Error Rate Calculation

- For a sentence pair

$$AER(i) = 1 - \frac{2 * |S_G \cap R_S|}{|S_G| + |R_S|}$$

- Calculate the error rate of a aligner
  - Based on the labeled data instead of the whole data

$$\varepsilon_l = \sum_{i \in D} w_l(i) * AER(i)$$

where

$w_l(i)$ **is the normalized weight of the i**th **sentence pair at the l**th **round**

Re-weight the Training Data

- Reweight each sentence pair in the training set
  - For each sentence pair, there may exists correct links and incorrect links as compared with the pseudo reference set
  - Calculate the weight of each sentence pair according to the correct and incorrect links

$$w_{l+1}(i) = w_l(i) * (k + (n-k) * \beta_l) / n$$

where $\beta_l = \varepsilon_l / (1 - \varepsilon_l)$

K **is the number of the error links**

n **is the total number of the links in the reference**

Final Ensemble

- Obtain the final ensemble according to the trained word aligners on each round

$$h_f(s) = \arg\max_t \sum_{l=1}^{L} (\log \frac{1}{\beta_l}) * WT_l(s,t) * \delta(h_l(s) = t)$$

where $$WT_l(s,t) = \frac{2 * count(s,t)}{\sum_{t'} count(s,t') + \sum_{s'} count(s',t)}$$

$WT_l(s,t)$ is the weight of each alignment pair (s,t) produced by the word aligner h_l

h_f is the final ensemble for word alignment

$\log \frac{1}{\beta_l}$ is the weight of the word aligner h_l

Evaluation

- Training set
  - Unlabeled data: 320,000 English-Chinese pairs
  - Labeled data: 30,000 English-Chinese pairs
- Held-out set
  - 1,500 sentence pairs
- Testing set
  - 1,000 bilingual English-Chinese sentence pairs
  - Totally 8,651 alignment links
  - 866 multi-word alignment links

Evaluation Metric

- Word alignment
  - Precision and Recall
  - Alignment Error Rate (AER)

$$Precision = \frac{|S_G \cap S_C|}{|S_G|} \qquad Recall = \frac{|S_G \cap S_C|}{|S_C|}$$

$$AER = 1 - \frac{2 * |S_G \cap S_C|}{|S_G| + |S_C|}$$

- Phrase-based machine translation
  - BLEU, NIST

Word Alignment Results

Method	Precsion	Recall	AER
Interpolation	0.7555	0.7084	0.2688
Supervised Boosting	0.7771	0.6757	0.2771
Unsupervised Boosting	0.8056	0.7070	0.2469
Semi-supervised Boosting	0.8175	0.7858	0.1987

Weights in Ensembles

- Two kinds of weights
  - Weights for the individual aligners
  - Weights for the individual alignment links

Method	Precision	Recall	AER
Baseline	0.7946	0.7775	0.2140
Our method	0.8175	0.7858	0.1987

Baseline: only use the first kind of weights

Our method: use the two kinds of weights

Translation Results

Method	NIST	BLEU
Interpolation	4.7929	0.1350
Supervised Boosting	4.4296	0.1151
Unsupervised Boosting	4.9045	0.1459
Semi-supervised Boosting	5.1729	0.1525

DEMO