IWSLT 2004 Workshop

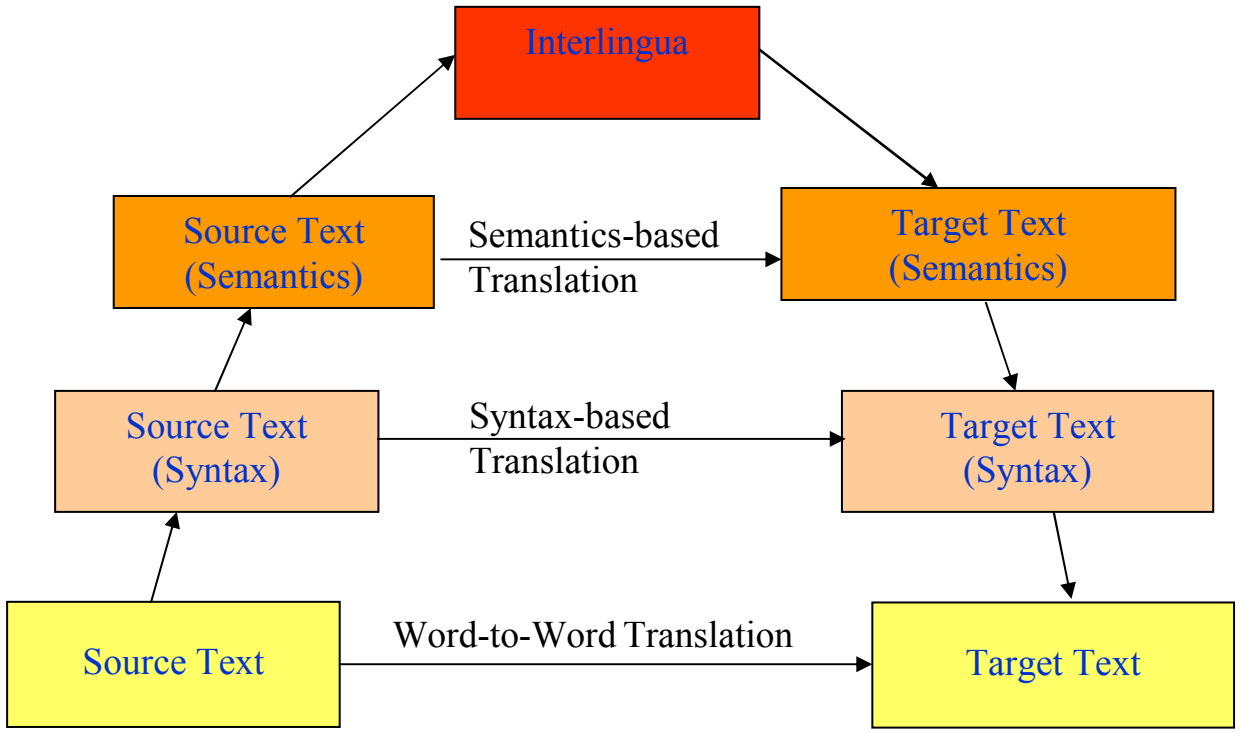
The ISL EDTRL System

Jürgen Reichert
University of Karlsruhe

Overview

1. Introduction
2. Basic Ideas of EDTRL
3. Training and Translating
4. Experiments and Results

1. Introduction



level of abstraction

Usage of Knowledge

- Knowledge-based Approaches
  - Grammar writing
  - Expert system
  - Frame-based Machine Translation
- Data-driven Approaches
  - Example-based Machine Translation
  - Statistical Machine Translation
  - Grammar learning

2. Basic Ideas of EDTRL

Combine the Approaches of

- Interlingua Systems
  (2n Modules, NLP, Paraphrases)
- Data-Driven Systems
  (only depends from corpus)
- Knowledge-based Approaches
  (further semantic and morphologic knowledge)

What do we get?

⇒Domain only depends from training corpus

⇒No handcrafted work

⇒Easily add new language

⇒Reduce Parallel Data Sparseness Problem

- Very few

  Spanish – Chinese

- Large amount of

  Chinese – English

  English – Spanish

-> use Chinese – English – Spanish

Enriched and Formalized English as Interlingua

1. Cascaded translation
2. Preserve translation alternatives
3. Confidence Measures / Probabilities
4. Standardized and Simplified English
5. Linguistically enriched English

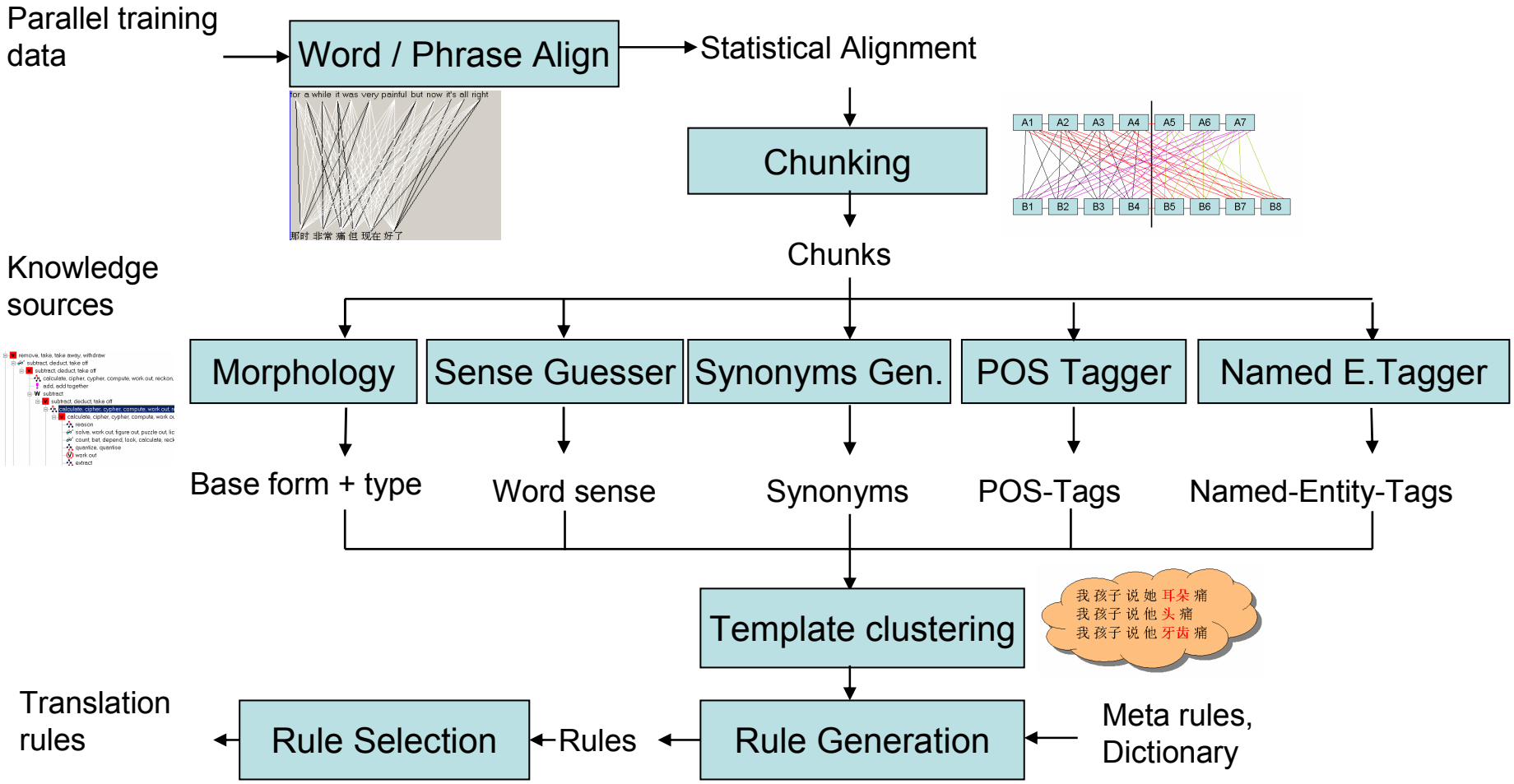
Formalizations & Enrichment

- Simplified English (common used alternative)
- Standardized word order (please give me … -> give me … please)
- Added Attributes
  - Morphological knowledge
  - Sense
  - Synonym Generator
  - Part-of-Speech Tags
  - Named Entity Tags
  - (sentence type, active/passive, politeness, domain, category, …)

Statistical Translation Rules

- Transfer knowledge from English to the foreign language using the statistical Alignment
- Error-Driven Learning (learn from errors)
- Interactive learning modus
- Translation error tracking and correction
- Small memory/time footprint (scalable)

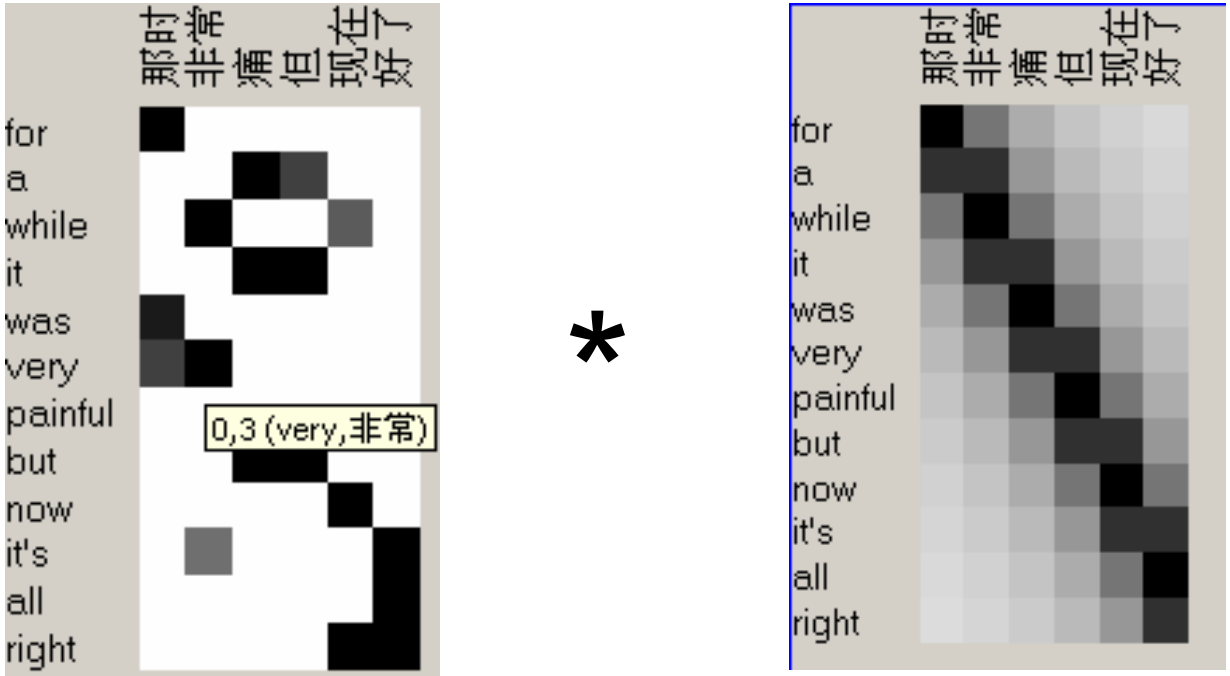
3. EDTRL Training



Weight functions for Alignment

1. Weight Position factor

$$\frac{1}{\left| WordPos1 - \left(WordPos2 \bullet \frac{\#Words1}{\#Words2} \right) \right|}$$

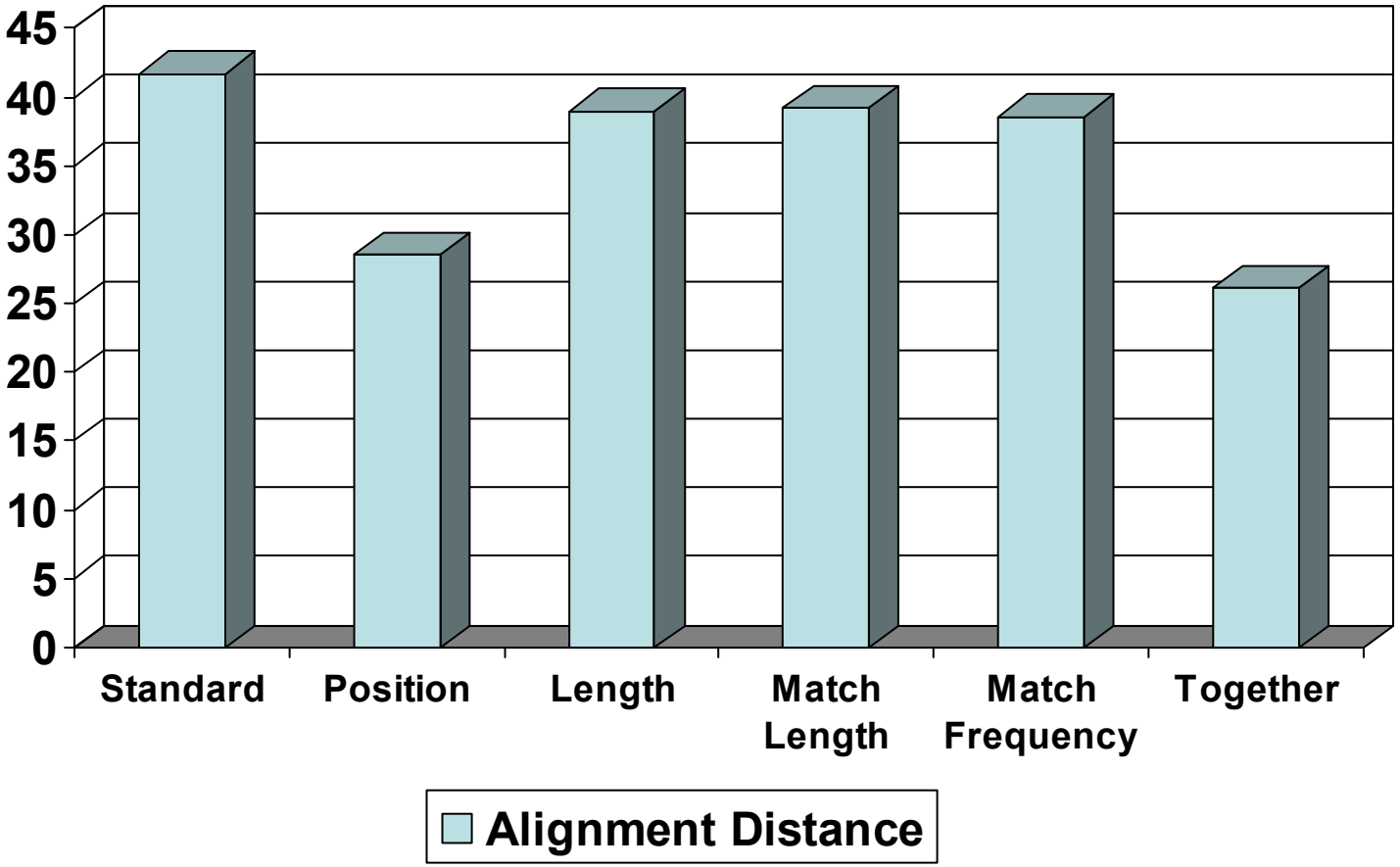


Weight functions for Alignment

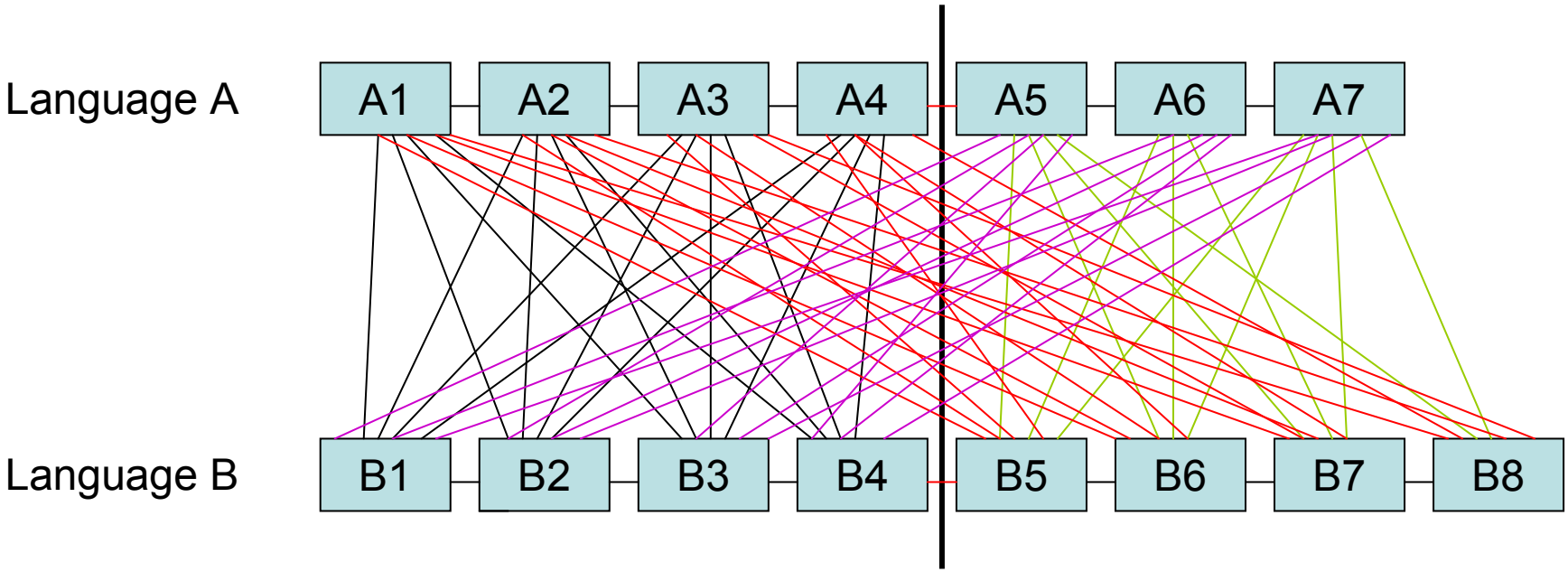
2. Length penalty $\frac{k}{\log(len)}$
3. Matching Length factor (prefer same length) $\frac{\#LenA+\#LenB}{2\cdot\max(\#LenA,\#LenB)}$
4. Frequency Weight (prefer alignment between words with similar frequency)

$$\frac{\#wordsA+\#wordsB}{2\cdot\max(\#wordsA,\#wordsB)}$$

Weight functions for Alignment



Chunking

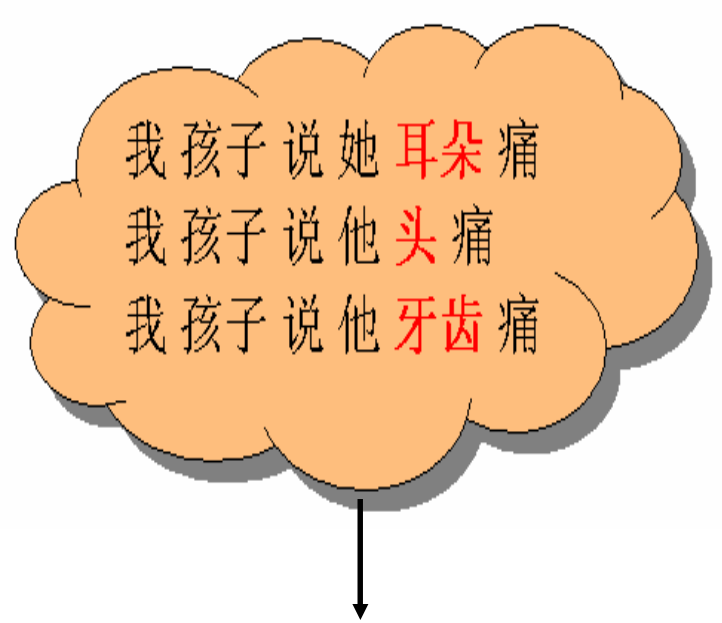


Align Dependence = (align weigths + align weights) / (align weigths + align weights)

LM Dependence = Sum $P(W_n|W_{n-1})$ / Sum $P(W_n|W_{n-1})$

Template generation

1. Cluster similar Sentence pairs
2. Generate Phrase Alignment
3. Build templates with classes for the different words



我孩子说她<0:bodypart>痛

Rule Generation/Selection

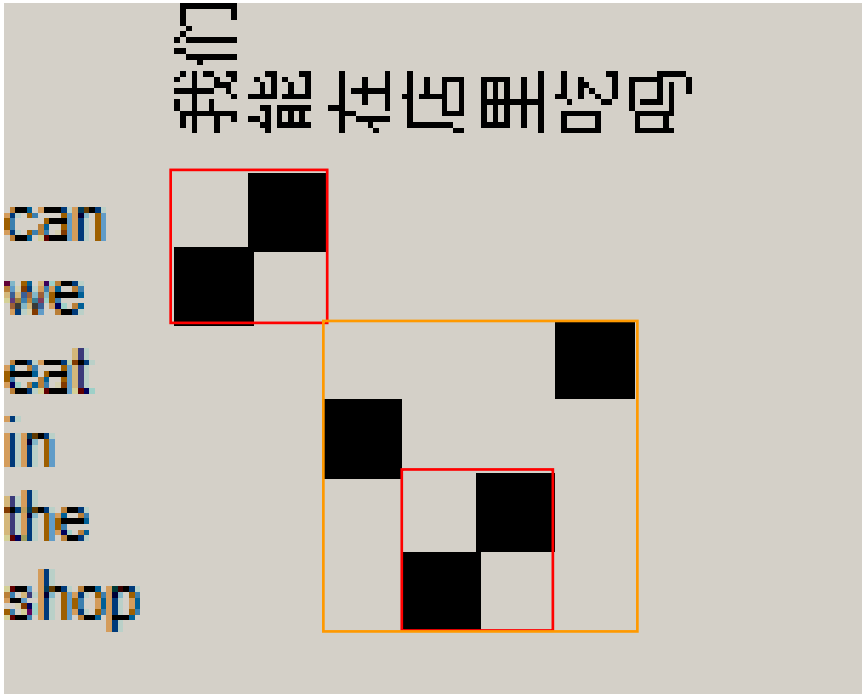
- Rules

  Cond1 | Cond2 | … → Templ1 | Templ2 | …
  Build form: Word, Phrase, Attribute Class

  Scores (Probabilities) for each Template
- Find ‘optimal’ rules
- Evaluate rule on verification set
- Using a class hierarchy
- Using meta-rules for the construction

Learning reorder rules

- Search reorders with a high alignment confidence
- Generalize or specialize the reorder rules by introducing classes and conditions



Translation

- Left to Right
- Find Matching rules -> probability
- Instantiate rules -> probability
- Beam-Search weighted by a trigram
- Pruning

Translation

A) Chinese -> IL (Tagged English):

我 想 我 从 某人 那 传染 上 感冒 了

我 从 某人 那 <1> 上 <2> 了

-> I've <1:VB> a <2:Disease> from someone

传染: infection <NN>0.3, transmission <NN>0.1, infect <VB>0.1, catch <VB>0.2

感冒: cold <Disease>0.3, rheum <Body Substance>0.2, to catch cold <Change>0.4

Instantiation => I think I've caught a cold from someone

Translation

B) IL (Tagged English) -> Chinese/Spanish :

I think I've caught<1:VB> a cold <2:Disease> from someone

I've <1:VB> a <2:Disease> from someone

-> 我 从 某人 那 <1> 上 <2> 了

catch <VB>: 捕捉 0.4, 逮 0.3, 传染 0.1 …

catch <NN>: 陷阱 0.1, …

cold <Temperature attribute>: 冷0.4 ， 凉 0.4

cold <Disease>: 感冒 1.0

Instantiation => 我 想 我 从 某人 那 捕捉 上 感冒 了

4. Experiments and Results

Preprocessing

- IWSLT
- New Segmentation (for Chinese)

Post processing

- a -> an
- removing duplicates
- Some verb form adaption

Database

	Training			Test
	# English Phrases	# Chinese Phrases	# Spanish Phrases	
BTEC	162314	162314	6027	506
Medical	7634	7634	7634	200
Tourism	2003	2003	2003	200
$\sum$	171951	171951	15664	

Experiment: mixing Domains

Train	Test	NIST-score
BTEC (162000)	BTEC (506)	4,7109
medical (6500)	medical (200)	2,8434
tourism (2000)	tourism (200)	3,1706
btec+medical+tourism	btec	4,7617
btec+medical+tourism	medical	2,8952
btec+medical+tourism	tourism	3,1735
btec+medical+tourism	btec (large system)	4,8383

Results

Systems	EDTRL	Systran
C → E	7,34	5,74
E → S	5,17	6,06
C → S	3,17	-
C → E → S	3,41	2,84
C → E_{IL} → S	3,69	-

IWSLT 2004 evaluation Chinese-English unrestricted

Method	Score	Rank (of 9)
fluency	2.93	6
adequacy	3.25	3
BLEU	0.27	5
GTM	0.66	4
NIST	7.50	2
PER	0.42	3

Conclusion

- The EDTRL System has a better performances than simple cascaded multiple MT systems.
- The use of formulized, enriched English as Interlingua can reduce the Parallel Data Sparseness Problem form many languages pairs
- Results from IWSLT 2004 evaluation campaign lie behind the best systems