Hybrid MT
Towards integrating modules and methods

Kurt Eberle
Term Paper

• 10 – 15 pages
1. What is the main idea of the paper presented?
2. What is the corresponding approach?
3. What are the main results?
   (1-3 relatively short, in own words)
4. What about the approach and findings when related to other approaches?
5. What are Pros and Cons?
The basic architectures RBMT & SMT
The basic architectures RBMT & SMT

RBMT

SMT

Alignment

Tokens

phrase S

abstract phrase S (Hiero, Treelet)

abstract phrase T

phrase T

More abstract representations

Syntactic representation

Universal representation

Interlingua systems

Transfer systems

Transducers

Sentence

Translation

Words
The basic architectures RBMT & SMT

RBMT

SMT

Alignment model
The basic architectures RBMT & SMT

RBMT

SMT

Alignment

Tokens

cover

phrase S

abstract phrase S (Hiero, Treelet)

Transfer systems

Transducers

Sentence

Translation

Words

Alignment model

Language model

Interlingua systems

Universal representation

More abstract representations

Syntactic representation

Tokens

reorder

phrase T

abstract phrase T
The basic architectures RBMT & SMT

RBMT

- Tokens
- Alignment
- abstract phrase S (Hiero, Treelet)

SMT

- Tokens
- Alignment
- abstract phrase T
- Language model
- Translation model
- Alignment model

Interlingua systems

Universal representation

More abstract representations

Syntactic representation

Translation model
The basic architectures RBMT & SMT

RBMT

SMT

Alignment

Tokens

cover

phrase S

abstract phrase S
(Hiero, Treelet)

Tokens

reorder

phrase T

abstract phrase T

Translation model

Alignment model

Language model

Decoder

Interlingua systems

Universal representation

More abstract representations

Syntactic representation

Translation model
Coupling RBMT systems & SMT Decoder

Decoder

EiseleFedermann et al
Coupling RBMT systems & Language Model
Coupling RBMT systems & Controller
Coupling analyses in RBMT

Selector
MS
Federmann
Pipeline (of RBMTs): Pivot translation

Wu Wang

Selection: best of n T2(T1(S))
Pipeline (of RBMTs): Corpus generation for SMT

Source parallel corpus

Target

Wu Wang

SMT
RBMT: Corpus generation for SMT

New Corpora from RBMT for SMT

Eisele et al, Dugast et al (Systran)

Input

Transducers

Sentence

Tokens

Alignment

phrase S

abstract phrase S

Source

parallel corpus

Target

Output

Transfsers

Translation

Words

Universal representation

Interlingua systems

Tokens

phrase T

abstract phrase T

RBMT

SMT
Pipeline: Preparation, MT, Postprocessing

Preparation

RBMT

SMT

Reorder input according to target (using parses and alignment)

+ by automatically extracting rewrite patterns: McCord Xia

Postprocessing

Tokens

Alignment

Translation

Words

Transfer

More abstract representations

Syntactic representation

Interlingua systems

Universal representation

Transfer systems

Sentence

Translation phrase S abstract phrase T

phrase T

Tokens
Pipeline: Preparation, MT, Postprocessing

RBMT
Preparation
SMT

Reorder input according to target (using parses and alignment)
* by automatically extracting rewrite patterns: McCord Xia
* by manually extracting rewrite patterns Collins et al

Translation
Alignment
Tokens

Transfer systems
Interlingua systems
Universal representation
More abstract representations
Syntactic representation
Words

Sentence

Preparation
Postprocessing

Transducers
Pipeline: Preparation, MT, Postprocessing

Preparation

RBMT

SMT

Transfer systems

Transducers

Sentence

Translation

Translation

Words

Alignment

Tokens

Tokens

phrase S

Interlingua systems

Universal representation

More abstract representations

Syntactic representation

RBMT output -> improved output

* by SMT
Simard et al
Dugast et al
Use and improve components and resources

improve RBMT bilingual lexicon using syntax annotated data
Culo, Bernth McCord

RBMT

SMT

Source parallel corpus Target
Use and improve components and resources

improve RBMT bilingual lexicon using PoS annotated phrases and GIZA++ alignment

Eisele Federmann et al
Use and improve components and resources

improve MT bilingual lexicon by WSD using Wordnet and Corpora

Vintar et al

RBMT

SMT

Source parallel corpus Target

Wordnet

phrase S phrase T

Tokens

Transfer

Transfer system

Transducers

Sentence

Translation

Words

Syntactic representation

Incr...
Use and improve components and resources

improve RBMT transfer using syntax annotated data and alignment statistics

Lavoie et al, Franz et al

RBMT

SMT

Source parallel corpus Target
Use and improve components and resources

RBMT transfer using phrase tables and vice versa

Sánchez-M.et al
Sánchez-C et al

Transducers
Sentence
Translation
Tokens
phrase S
Alignment
phrase T

Transfer system
Transfer
Syntactic representation
Words
Use and improve components and resources

improve alignment using syntax annotated data and bilingual lexicon (anchors)

Ma Way

RBMT

SMT

Source parallel corpus Target
Use and improve components and resources

improve alignment using several SMT aligners and RB aligner

RBMT

SMT
Use and improve components and resources

improve phrases using RBMT's transfer knowledge

Eisele Federmann
Use and improve components and resources

improve phrases using RBMT for corpus generation

Eisele Federmann

RBMT

SMT

Tokens

Alignment

Tokens

phrase S

phrase T

Source

parallel corpus

Target
Hierarchical Phrase-based MT

Hierarchical PBMT

Cbiang

abstract phrase S \rightleftharpoons \text{Alignment} \rightarrow \text{abstract phrase T}

SMT

SCFG

Source parallel corpus Target
Use and improve components and resources

Treelet translation

Cherry et al

RBMT

SMT

Source
parallel corpus
Target

Transfer systems
Transducers
Sentence
Translation
Words

Tokens
Alignment
Tokens

treelet S
phrase T

More abstract representations
Syntactic representation

Source
parallel corpus
Target
Establish and Use Relations

Chen et al

improve transfer using hierarchical phrases

RBMT

Preparation

SMT

Postprocessing

Tokens

Alignment

Tokens

Sentence

Translation

Words

Transducers

Transfer

Syntactic representation

Transfer system

Integrate representations

Hierarchical phrase S

Hierarchical phrase T
Measures

• Bleu, Rouge, Meteor, MER, ...
• Different pros and cons
• Different purposes
Findings

• Many ways of hybridization
• Quality estimation can be dependend on measure used

• Typically, RBMT components are smaller than corr. SMT comps.
• Avoid manual working out
  • Learn Grammar and Parser from data
  • Learn Lexicon from data and syntactic analysis
  • ....
Outlook

N.B.: meaning of a rule ....

\[
[[ \text{NP}_{[\text{Gen} = m]} \leftarrow \text{DET}_{[\text{Gen} = m]} \text{NP}_{[\text{Gen} = m]} ]] \\
= \{ \text{der Mann, der Tisch, der Rahmen,}...\}
\]

\[
[[ \text{NP}_{[\text{Gen} = X]} \leftarrow \text{DET}_{[\text{Gen} = X]} \text{NP}_{[\text{Gen} = X]} ]] \\
= \text{Det N sequences which are grammatically correct!}
\]

The meaning of a rule is the set of things that satisfy the rule!

( Corpus ) Extension → Abstraction → Rule

Rule → Application → Extension ( Corpus )