Generalized EBMT

March 9, 2006

Jaedy Kim

Advised by

Ralf Brown
Peter Jansen

Language Technologies Institute

Overview

- Introduction
- CMU G-EBMT System
  - Matching
  - Generalization
  - Alignment
  - Decoding
  - Applications

Overview

- Introduction
  - What is EBMT?
  - An example run
  - Types
  - Components
  - Advantage
- CMU G-EBMT System

What is EBMT?

- Basic Philosophy
  - "man does not translate a simple sentence by doing deep linguistic analysis, rather, man does translation, first, by properly decomposing an input sentence into certain fragmental phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference." -- Nagao (1984)

An example run (1)

Translation Examples

1. Yesterday 200 delegates met begind closed doors to discuss the new tax code.
   Gestern trafen sich 200 Abgeordnete hinter verschlossenen Tueren,
   um ueber die neuen Steuergesetze zu verhandeln.

2. Gennifer Flowers is said to have haad an affair with President Clinton for many years.
   Gennifer Flowers hat angeblich jahrelang eine Affaere mit Praesident Clinton gehabt.

Yesterday 200 deligates met with President Cliton

An example run (2)

Translation Examples

1. Yesterday 200 delegates met begind closed doors to discuss the new tax code.
   Gestern trafen sich 200 Abgeordnete hinter verschlossenen Tueren,
   um ueber die neuen Steuergesetze zu verhandeln.

2. Gennifer Flowers is said to have haad an affair with President Clinton for many years.
   Gennifer Flowers hat angeblich jahrelang eine Affaere mit Praesident Clinton gehabt.

Yesterday 200 deligates met with President Cliton

Gestern trafen sich 200 Abgeordnete

An example run (3)

Translation Examples

1. Yesterday 200 delegates met begind closed doors to discuss the new tax code.
   Gestern trafen sich 200 Abgeordnete hinter verschlossenen Tueren,
   um ueber die neuen Steuergesetze zu verhandeln.

2. Gennifer Flowers is said to have haad an affair with President Clinton for many years.
   Gennifer Flowers hat angeblich jahrelang eine Affaere mit Praesident Clinton gehabt.

Yesterday 200 deligates met with President Cliton

Gestern trafen sich 200 Abgeordnete mit Praesident Clinton

Where is EBMT?

- Rule-based systems
- Interlingual systems
- Corpus-based systems
  - Statistical Machine Translation
  - Translation Memory
  - Example-Based Machine Translation

EBMT vs Rule-based systems

- EBMT may include a component to automatically learn translation rules
  - Extracting bilingual terminology
  - Finding equivalence classes among words
  - Inducing morphology rules
  - Inducing grammar rules

EBMT vs TM

- TM
  - A tool to aid a human translator
    - Most useful when translating revised versions of previously-translated documents
- EBMT
  - Generalized TM
    - Find the nearest matching sentence and determine how to transfer any remaining differnces – may need knowledge
    - Find the largest exact matches of portions of the input to be translated, and combine the pieces later

EBMT vs SMT

- SMT
  - Consists of one or more mathematical models:
    - Translation probabilities
    - Word re-ordering probabilities
    - Output language model
  - Combines word and short n-gram translations well
  - Cannot readily exploit long pre-translated phrases.
- EBMT
  - Exploit long translated phrases
  - Doesn not combine phrasal translations well

Types of EBMT systems

- Lexical (shallow)
- Morphological / part-of-speech analysis (less shallow)
- Parse tree-based (deep)

Resources

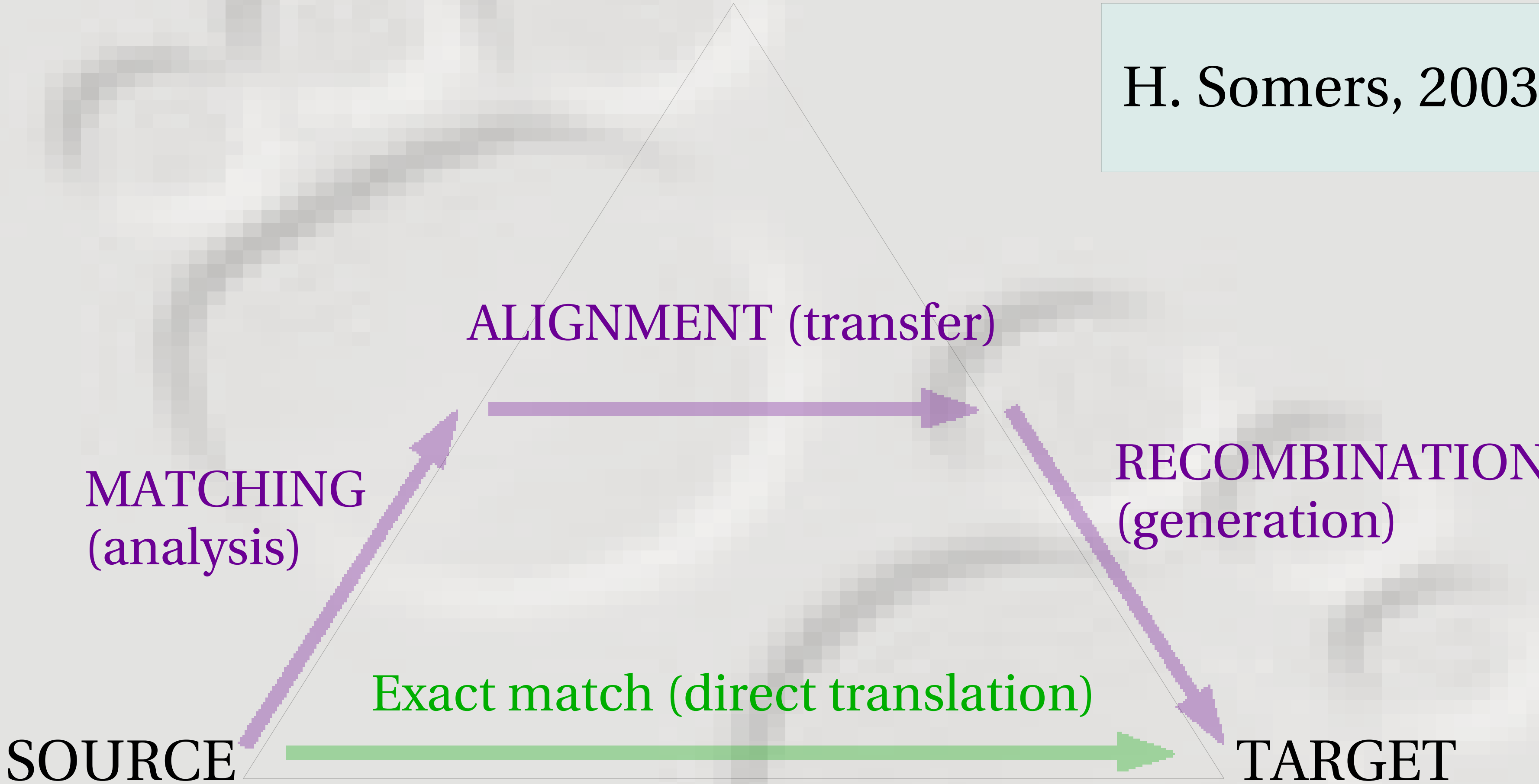
- Types of data/knowledge required by EBMT systems
  - Parallel text
  - Bilingual dictionary
  - Thesaurus for computing semantic similarity
  - Syntactic parser, dependency parser, etc.
- World Wide Web
  - As a source of parallel text
  - As a means of validating translations

Main components

- Matching
  - Finds real examples for input
- Alignment
  - Identifies corresponding translation fragments
- Recombination
  - Recombines translation fragments into target text

EBMT "Pyramid"

H. Somers, 2003



Matching

- String matching
  - IR style matching
  - Cheap but needs a lot of data
- Meaning matching
  - Based on semantic similarity using thesaurus
  - Needs linguistic knowledge
- Structure matching
  - Tree edit distance
  - Needs linguistic knowledge

Alignment

- Once close examples to the input are found
  - Identifies the translation fragments of matched source fragments
    - Word level
    - Template level
    - Tree level

Recombination

- Once target translation fragments are found
  - We combine them to get a target text
  - The resultant partial translations may not 'fit together' properly (boundary friction)
    - Word level alignment may have included extraneous words or missed a necessary word
    - One or more fragments may have the wrong case, number, etc.
    - Fragments may not show the correct agreement with each other – His face was a / open book

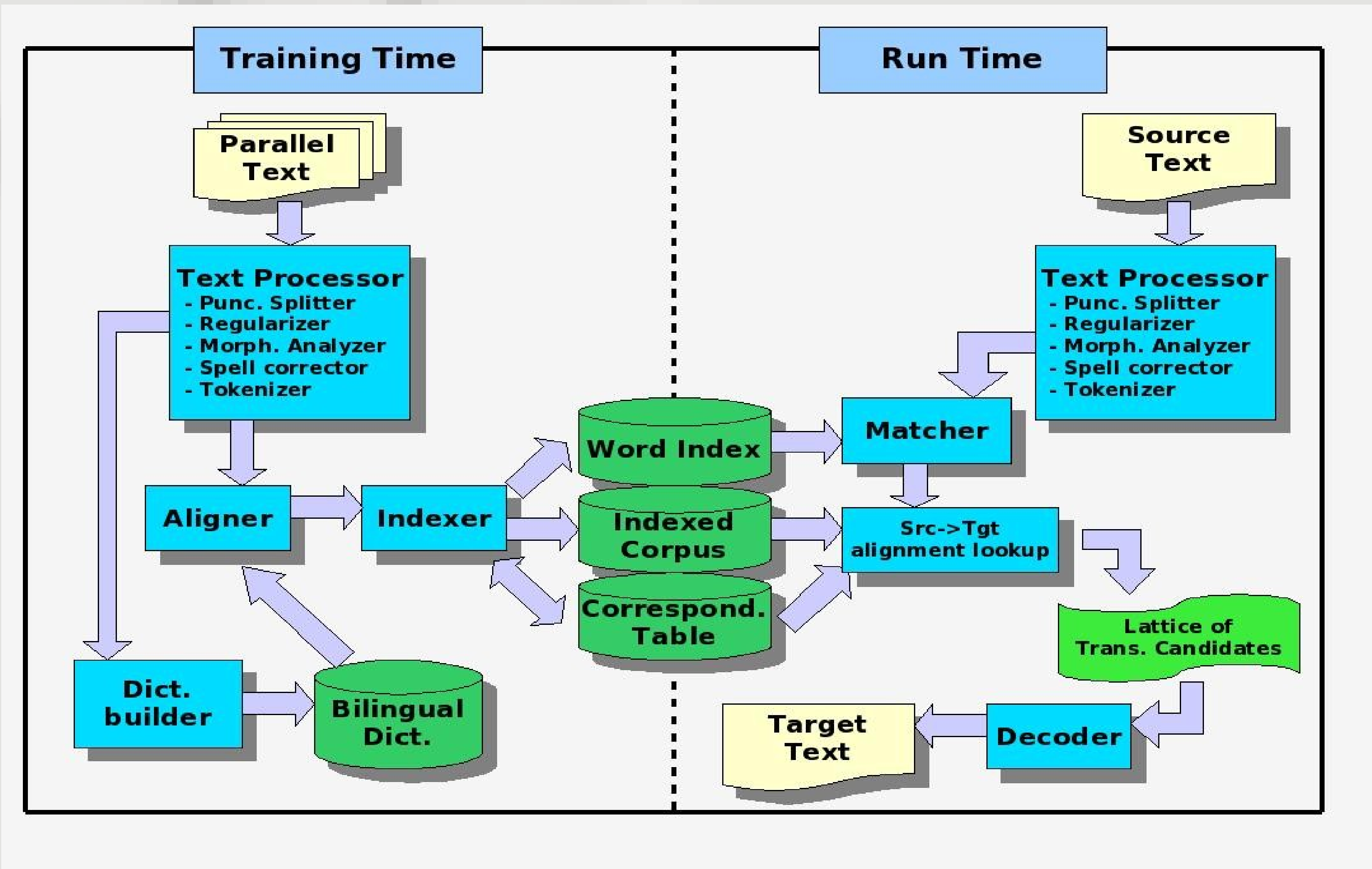
When is EBMT useful?

- Translation rule is difficult to formulate
- Translation of idioms/difficult linguistic phenomena (N1 no N2)
- Translation cannot be made by a compositional way using target words (different expressions)
- When sentence to be translated has a close match in the database

Overview

- Introduction
- CMU G-EBMT System
  - Matching
  - Generalization
  - Alignment
  - Decoding
  - Applications

Diagram



Training Time
Run Time
Parallel Text
Text Processor
- Punc. Splitter
- Regularizer
- Morph. Analyzer
- Spell corrector
- Tokenizer
Aligner
Indexer
Dict. builder
Bilingual Dict.
Word Index
Indexed Corpus
Correspond. Table
Source Text
Text Processor
- Punc. Splitter
- Regularizer
- Morph. Analyzer
- Spell corrector
- Tokenizer
Matcher
Src->Tgt alignment lookup
Lattice of Trans. Candidates
Decoder
Target Text

Matching (1)

- Lexical Matching : shallow processing
  - String match of surface forms
    - Advantage: little or no need for linguistic knowledge
    - Disadvantage: needs large amount of training data
  - Convert text into templates, and string match
    - Advantage: requires less training text
    - Problem: how to produce good-quality general templates?

Matching (2)

- Inexact Matching
  - Allows matches where not all words are matched
    - Allowance for a one-word gap when that word has
      - Only one translation listed in the dictionary
      - Its most-common translation occurring more than twice as frequently as the next translation
  - Helpful on limited training data but didn't improve quality when more data was available

Generalization

- Reduces the data need
- Approaches
  - Manually-generated equivalence classes
    - A machine-readable dictionary with part-of-speech
  - Automatically-extracted equivalence classes
    - Word-level clustering
  - Transfer-rule induction

Transfer-Rule Induction

- Assumption : when two sentence pairs in the corpus have some part in common but differ in some other part, the similar and dissimilar parts each correspond to some coherent constituent.
- Constituents : the groupings which can be used interchangeabley
- Searches for pairs of training instances where the source-language halves show the pattern

$$S_1 \; D \; S_2$$

The Induction Process

1. read the corpus into memory, creating a rough bitext mapping for each bilingual sentence pair
2. sort the corpus alphabetically by source-language sentence
3. for each F, find all sequences of sentence pairs which share the same first F words in the source language
4. for each sequence, create a subcorpus and:
    (a) sort the subcorpus alphabetically by **reversed** source-language sentence
    (b) for each L, find all sequences of sentence pairs which share the same last L words in the source language
    (c) for each sequence, create another subcorpus and:
        i. perform a pairwise comparison between sentence pairs, adding the differences to a new equivalence class and to the corpus. The bitext map is used to discard those differences which do not appear to match between source- and target-language sentences.
        ii. if sufficiently long, add the common initial/final strings to the corpus
5. apply the learned rewriting rules to the corpus, except to sentence pairs where doing so would generate a single token
6. repeat steps 2 through 5 until no more new equivalence classes are added or the number of iterations reaches a preset maximum

Example (1)

Sort sentences

We are watching agricultural chemicals .
Nous regardons les produits chimiques agricoles .
We are watching energy supplies .
Nous regardons les approvisionnements en energie .
We are watching equipment supplies .
Nous regardons les approvisionnements en materiel .
We are watching fertilizer supplies .
Nous regardons les approvisionnements en engrais .
We are watching steel production .
Nous regardons la production de acier .

Example (2)

Find differences

We are watching agricultural chemicals .
Nous regardons les produits chimiques agricoles .
We are watching steel production .
Nous regardons la production de acier .
We are watching energy supplies .
Nous regardons les approvisionnements en energie .
We are watching equipment supplies .
Nous regardons les approvisionnements en materiel .
We are watching fertilizer supplies .
Nous regardons les approvisionnements en engrais .

Example (3)

Replace entries with their class label

We are watching agricultural chemicals .
Nous regardons les produits chimiques agricoles .
We are watching steel production .
Nous regardons la production de acier .
We are watching <CL_0> supplies .
Nous regardons les approvisionnements en <CL_0> .

Make an equivalence class

<CL_0>:
"energy" = "energie"
"equipment" = "materiel"
"fertilizer" = "engrais"

Example (4)

Sorted by reverse word order again

We are watching agricultural chemicals .
Nous regardons les produits chimiques agricoles .

We are watching steel production .
Nous regardons la production de acier .

We are watching <CL_0> supplies .
Nous regardons les approvisionnements en <CL_0> .

Make an equivalence class

<CL_0>:
"energy" = "energie"
"equipment" = "materiel"
"fertilizer" = "engrais"

Example (5)

Replace entries with their class label

We are watching <CL_1> .

Nous regardons <CL_1> .

Make an equivalence class

<CL_0>:
"energy" = "energie"
"equipment" = "materiel"
"fertilizer" = "engrais"

<CL_1>:
"<CL_0> supplies" = "les approvisionnements en <CL_0>"
"agricultural chemicals" = "les produits chimiques agricoles"
"steel production" = "la production de acier"

Word Clustering

- For each source/target word pair, make a pseudo document
  - The word pair is picked in each sentence only when there are no translation alternatives
  - put immediately surrounding words of the source word into its pseudo document
- Word clustering became document clustering
- Word clustering can be done on a corpus generalized by grammar induction

Pseudo Document example

- Monolingual example

In the corpus

The **next** generation of mobile phone technology.
She was the **next** to appear.
He loved his dog **next** to his own sons.
Next time I come, I'll bring the children to you.

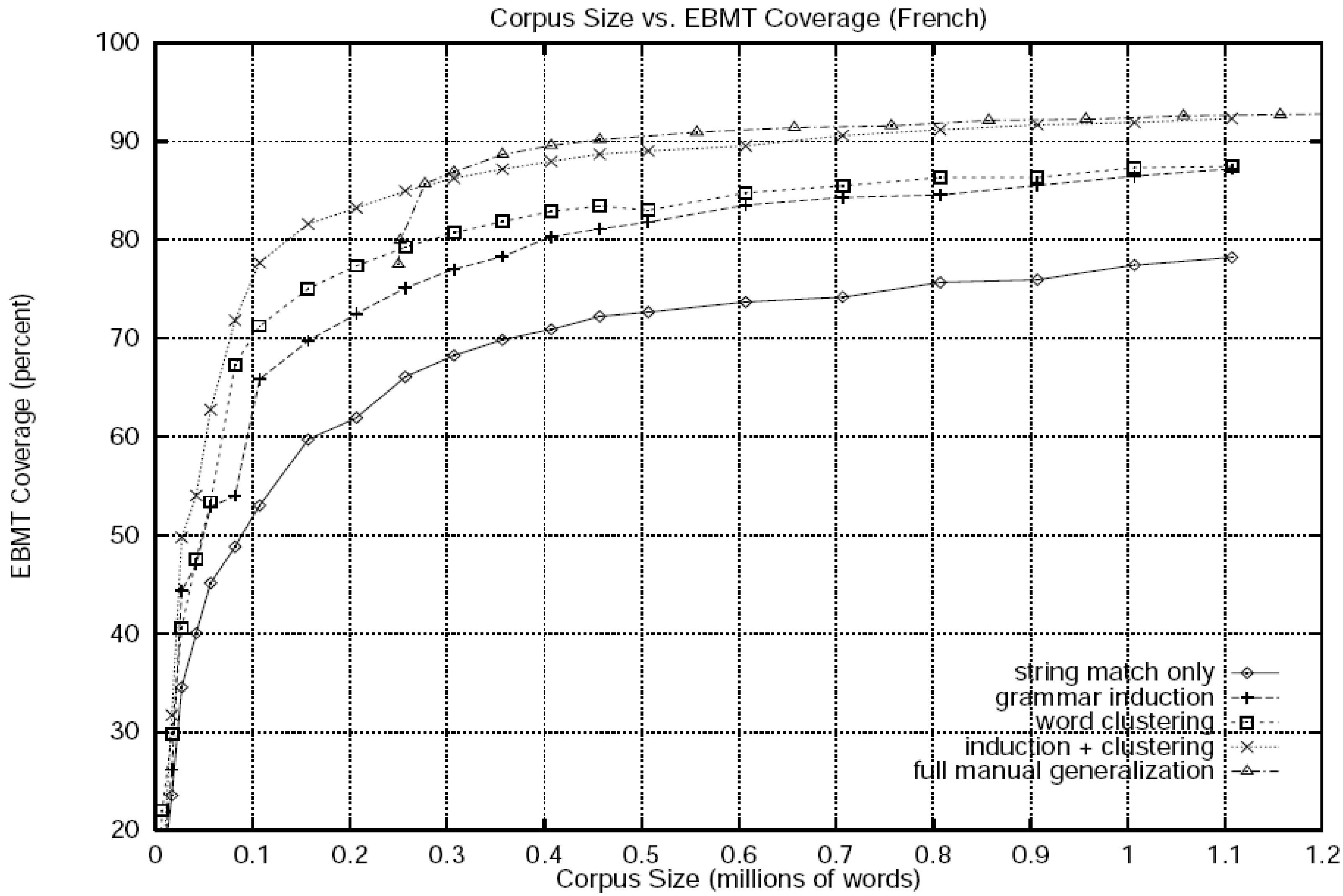
Pseudo document: next

The **next** generation of mobile
She was the **next** to appear .
loved his dog **next** to his own
Next time I come

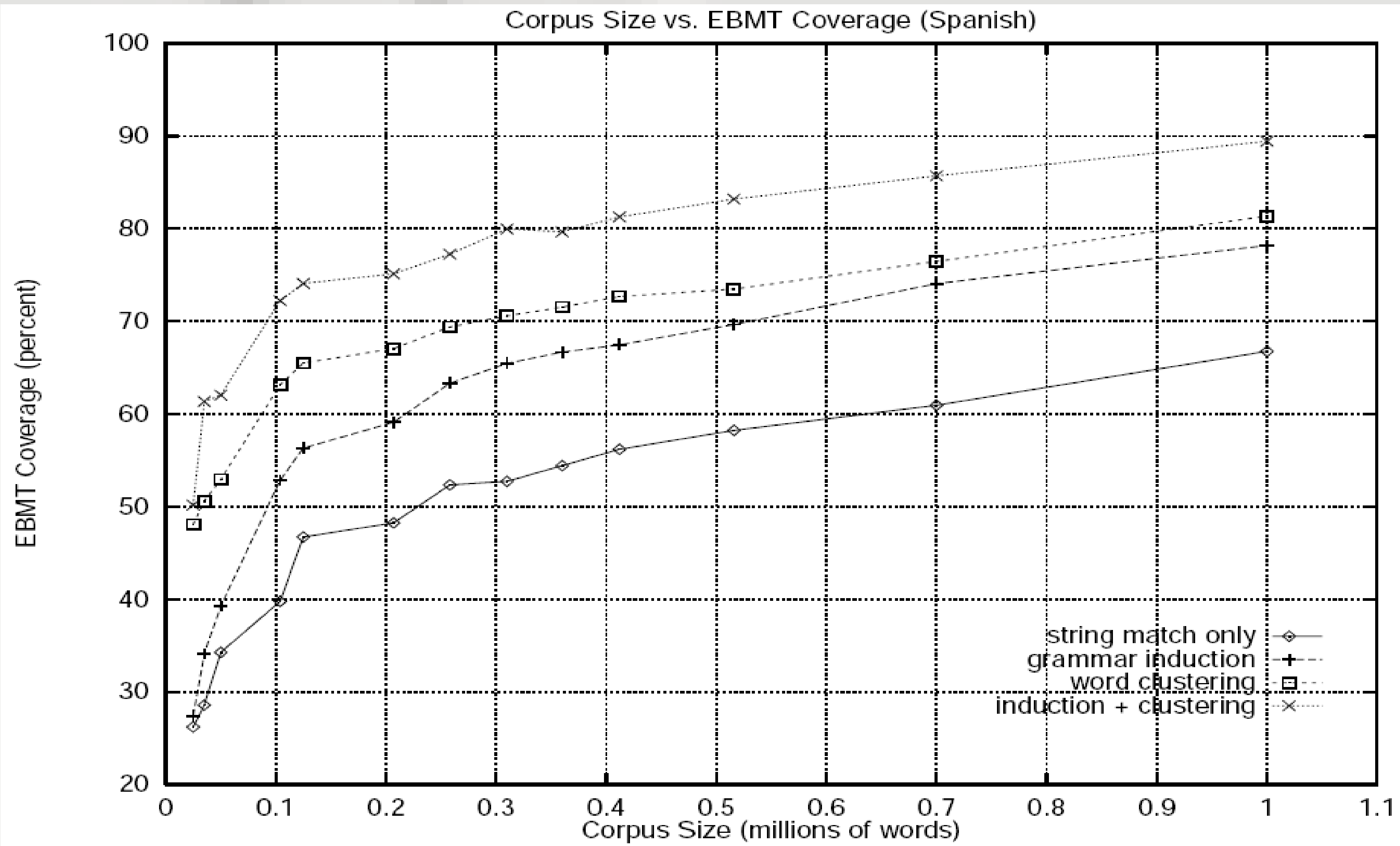
Example clusters

Cluster	French	English
507	NE	NOT
	NE	NO
524	SONT	ARE
	FURENT	WERE
568	CETTE	THIS
	CETTE	THAT
575	PROCHAIN	NEXT
	DERNIER	LAST
659	UNE	THE
	UNE	AN
	UNE	A
	UN	THE
	UN	A
	LE	THE
	LE	OF
	LE	IT
	LE	IN
	LE	A
	LA	THE
	LA	OF
	LA	IN

Results (1)


Corpus Size vs. EBMT Coverage (French)
EBMT Coverage (percent)
Corpus Size (millions of words)
100
90
80
70
60
50
40
30
20
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
string match only
grammar induction
word clustering
induction + clustering
full manual generalization

Results (2)

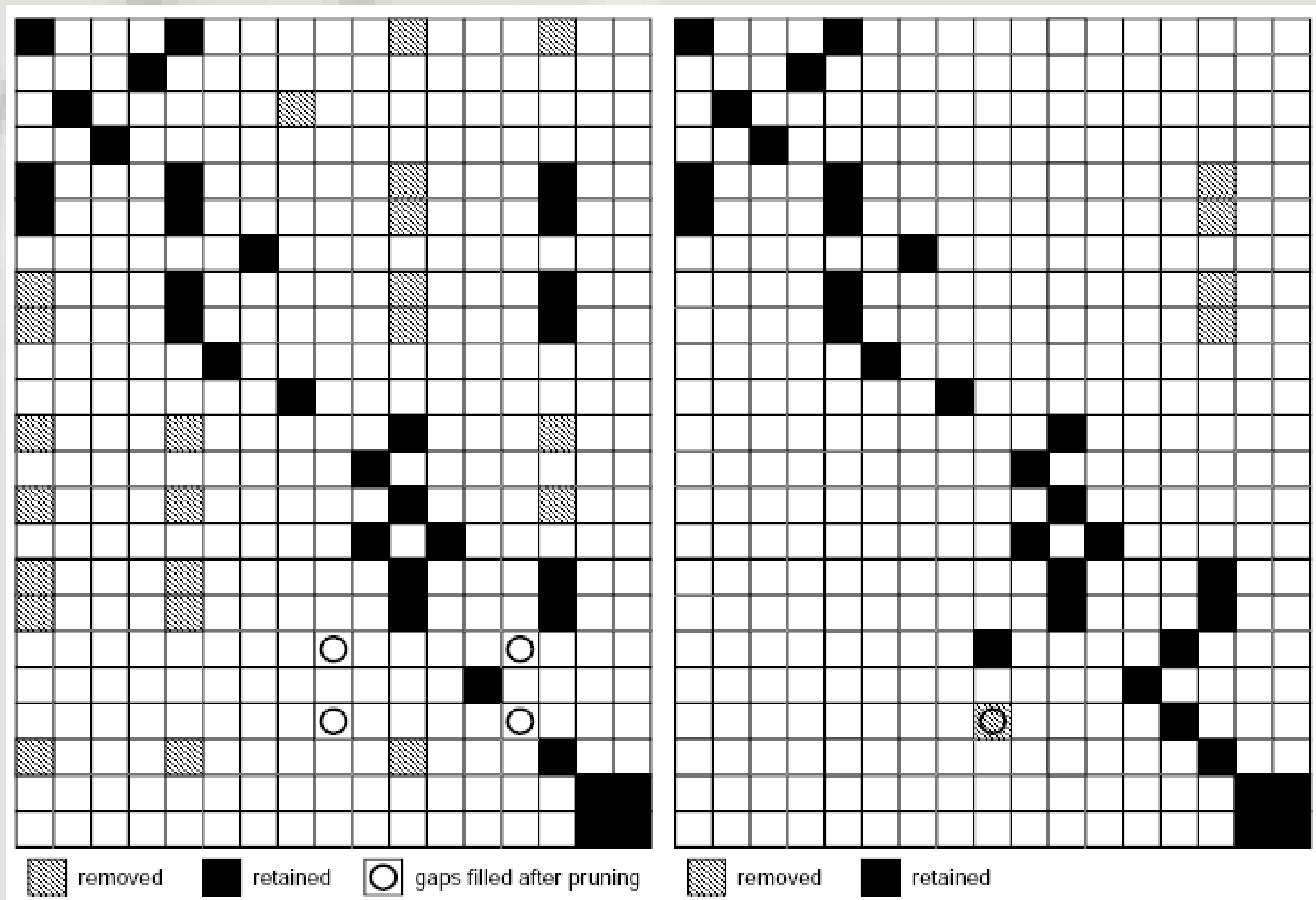


Corpus Size vs. EBMT Coverage (Spanish)
EBMT Coverage (percent)
Corpus Size (millions of words)
100
90
80
70
60
50
40
30
20
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
string match only
grammar induction
word clustering
induction + clustering

Word-level alignment

- To perform word-level alignment, it uses the translations in a bilingual dictionary along with heuristic scoring functions such as
  - Common location in sentence
    - Start, end of a sentence, positions of puncts
  - Difference in word length
  - Words known to translate as empty string

Correspondence Table



Symmetric Probabilistic Alignment

- Goal: given a sentence pair and a source fragment, find the most corresponding target fragment in terms of bi-directional translation probability
- Sub-sentential alignment
- Based on bidirectional probabilistic dictionary
- On-going research
- $$\bar{t} = argmax_t \left(\begin{array}{l} \left(\prod_{p=1}^{k}\left(max\left(max_{q=1}^{l} p\left(t_{j+q} \middle| s_{i+p}\right), \epsilon\right)\right)\right)^{\frac{1}{k}} \\ \times\left(\prod_{q=1}^{l}\left(max\left(max_{p=1}^{k} p\left(s_{i+p} \middle| t_{j+q}\right), \epsilon\right)\right)\right)^{\frac{1}{l}} \end{array} \right)$$

Decoding

- How to better combine phrasal translations?
- Traditional method
  - Required the source fragments to have no overlap
- Current method
  - Maximal left-overlap compositional MT
    - When both the source and target of two adjacent fragments overlap, then there is an increased likelihood their combination is an accurate translation

Example

Input:	Je	doute	qu'il	soit	nécessaire	de	lancer	une	enquête	complète	pour	l'instant.
Fragment												
1	Je	doute	qu'il									
	I do not think it is											
2	Je	doute	qu'il	soit								
	I doubt whether that will be											
3			qu'il	soit	nécessaire	de						
			not think it is necessary to									
4					nécessaire	de	lancer					
					necessary to start							
5								une	enquête	complète		
								a full investigation				
6											pour	l'instant.
											for the moment.	

Human reference translation:

"I do not think it is necessary to launch a full inquiry at this time."

Standard MEMT selection combines fragments 2, 4, 5, and 6, to produce the output:

"I doubt whether *that* will be necessary to start a full investigation for the moment."

MEMT selection with overlap combines fragments 1, 3, 4, 5, and 6, to produce the output:

"I do not think it is necessary to start a full investigation for the moment."

Finding the best path

- All translation candidates are placed into a common lattice, and each fragment is weighted by
  - The quality score assigned by the engine
  - The bonus factor given for longer fragments
  - The penalty factor for length mismatches between source and target halves
- Multi-level beam search guided by
  - The fragment weights
  - A target-language trigram language model

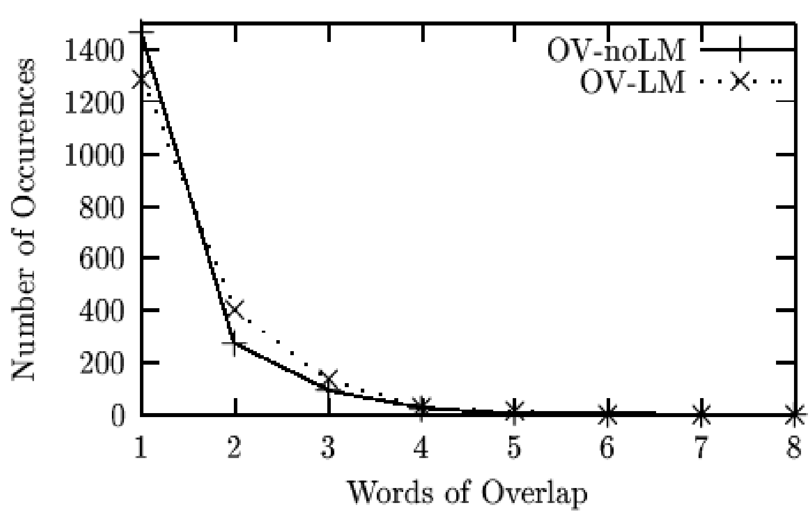
Multi-level beam search

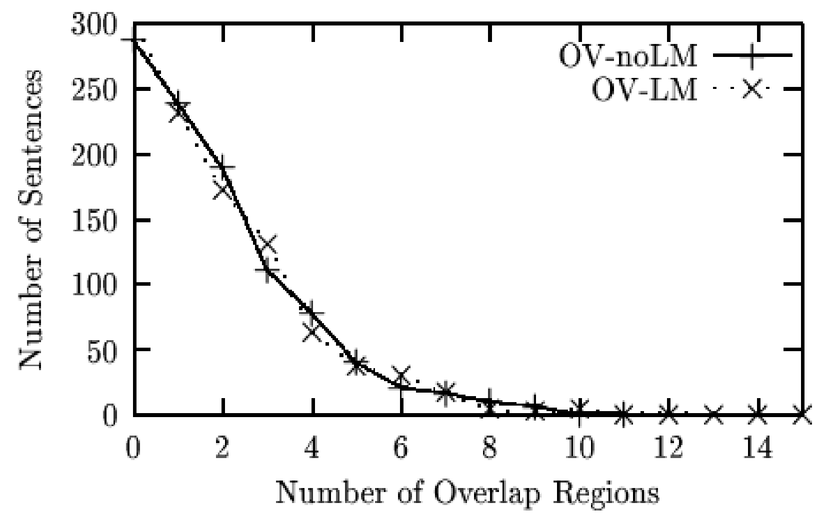
- A separate priority queue for each word position in the input text
- New nodes are added to the priority queues corresponding to the last input word covered by each extended path
- The priority queues are pruned by removing
  - The lower-scoring of duplicate nodes
  - The lowest-scoring node
- Overlapping fragment bonus to the path

Results

Training Set	System	Mean BLEU	St. Dev. BLEU	Mean NIST	St. Dev. NIST
Hansard 100k pairs	noOV-noLM	0.1412	0.0248	4.3017	0.3914
	noOV-LM	*0.1549**	0.0302	4.3591	0.3771
	OV-noLM	*0.1553**	0.0293	***4.4433****	0.4121
	OV-LM	***0.1707****+	0.0359	***4.4578****+	0.3942

Table 1: A performance summary of the described methods. Starred and plus-marked results are highly significant against the noOV-noLM and noOV-LM systems, respectively, according to the two-sided t-test ($p \leq 0.002$). Italicized and boldface results are significant against noOV-noLM and noOV-LM according to the sign test ($p \leq 0.05$).





Applications

- Text translation
  - Pangloss, Mega-RADD, Milli-RADD, AVENUE
- Speech translation systems
  - DIPLOMAT, TONGUES
- Cross-language information retrieval
  - MUCHMORE
- Topic tracking

Current/Future Work

- Generalization
- Alignment
- Morphology

References

- Ralf's tutorial on EBMT at AMTA-2002 (http://www-2.cs.cmu.edu/~ralf/ebmt/AMTA02-tutorial.pdf)
- Kevin Duh's Example-based Machine Translation: A Survey (http://ssli.ee.washington.edu/people/duh/projects/ebmt.pdf)
- Ralf D. Brown. "Transfer-Rule Induction for Example-Based Translation". In Proceedings of the MT Summit VIII Workshop on Example-Based Machine Translation, p. 1-11. Santiago de Compostela, Spain, 18 September 2001.
- Ralf D. Brown, Rebecca Hutchinson, Paul N. Bennett, Jaime G. Carbonell, Peter Jansen. "Reducing Boundary Friction Using Translation-Fragment Overlap", in Proceedings of the Ninth Machine Translation Summit, New Orleans, USA, September 2003, pp. 24-31.

Thank You!

Questions?

N1 no N2

- no” の is an adnominal particle
- Variants: “deno” での, “madeno” までの, etc.
- “Noun <u>no</u> Noun” => “Noun <u>of</u> Noun”

Youka **<u>no</u>** gogo	The afternoon **<u>of</u>** the 8th
Kaigi **<u>no</u>** mokuteki	The objective **<u>of</u>** the conference
Kaigi **<u>no</u>** sankaryou	The application fee **<u>for</u>** the conference
Kyouto **<u>deno</u>** kaigi	The conference **<u>in</u>** Kyoto
Isshukan **<u>no</u>** kyuka	A week**<u>’s</u>** holiday
Mittsu **<u>no</u>** hoteru	Three hotels