Improved Statistical Machine Translation for Resource-Poor Languages Using Related Resource-Rich Languages

Preslav Nakov

National University of Singapore

(joint work with Hwee Tou Ng)

EMNLP'2009

Overview

Overview

- Statistical Machine Translation (SMT) systems
  - Need large sentence-aligned bilingual corpora (bi-texts).
- Problem
  - Such large bi-texts do not exist for most languages, and building them is hard.
- Our (Partial) Solution
  - Use bi-texts for a resource-rich language to build a better SMT system for a related resource-poor language.

Introduction

Building an SMT System for a New Language Pair

- In theory: only requires few hours/days
- In practice: large bi-texts are needed
  - Only available for
    - Arabic
    - Chinese
    - the official languages of the EU
    - some other languages
  - However, most of the 6,500+ world languages remain resource-poor from an SMT viewpoint.

This number is even more striking if we consider language pairs.

Even resource-rich language pairs are resource-poor in most domains.

Building a Bi-text for SMT

- Small bi-texts
  - Relatively easy to build
- Large bi-texts
  - Hard to get, e.g., because of copyright
  - Sources: parliament debates and legislation
    - national: Canada, Hong Kong
    - international
      - United Nations
      - European Union: *Europarl*, *Acquis*

Becoming an official language of the EU is an easy recipe for getting rich in bi-texts quickly.

Not all languages are so "lucky", but many can still benefit.

Idea: Use Related Languages

- Idea
  - Use bi-texts for a resource-rich language to build a better SMT system for a related resource-poor language.

Some Languages That Could Benefit

- Related nonEU–EU language pairs
  - Norwegian – Danish
  - Macedonian[1] – Bulgarian
  - (future: Croatia is due to join the EU soon)
    - Serbian[2], Bosnian[2], Montenegrin[2] (related to Croatian[2])
- Related EU languages
  - Czech – Slovak
  - Spanish – Portuguese
- Related languages outside Europe
  - Malay - Indonesian

We will explore these pairs.

- Some notes:
  - [1] Macedonian is not recognized by Bulgaria and Greece
  - [2] Serbian, Bosnian, Montenegrin and Croatian were Serbo-Croatian until 1991; Croatia is due to join the EU soon.

Motivation

- Related languages have
  - overlapping vocabulary (cognates)
    - e.g., *casa* (‘*house’)* in Spanish and Portuguese
  - similar
    - word order
    - syntax

Example: Malay & Indonesian

- Malay–Indonesian

~50% exact word overlap

The actual overlap is even higher.

  - Malay
    - **Semua** manusia **dilahirkan** bebas **dan** samarata dari segi kemuliaan **dan hak-hak**.

      Mereka mempunyai pemikiran **dan** perasaan **hati dan** hendaklah bertindak di antara **satu sama lain** dengan **semangat persaudaraan**.
  - Indonesian
    - **Semua** orang **dilahirkan** merdeka **dan** mempunyai martabat **dan hak-hak** yang sama.

      Mereka dikaruniai akal **dan hati** nurani **dan** hendaknya bergaul **satu sama lain** dalam **semangat persaudaraan**.

(from Article 1 of the *Universal Declaration of Human Rights*)

Example: Spanish & Portuguese

17% exact word overlap

- Spanish–Portuguese
  - Spanish
    - **Todos** los **seres humanos** nacen libres **e** iguales en dignidad y derechos y, **dotados** como están **de** razón y conciencia, deben comportarse fraternalmente los unos con los otros.
  - Portuguese
    - **Todos** os **seres humanos** nascem livres **e** iguais em dignidade e em direitos. **Dotados de** razão e de consciência, devem agir uns para com os outros em espírito de fraternidade.

Example: Spanish & Portuguese (cont.)

17% exact word overlap
67% approx. word overlap

The actual overlap is even higher.

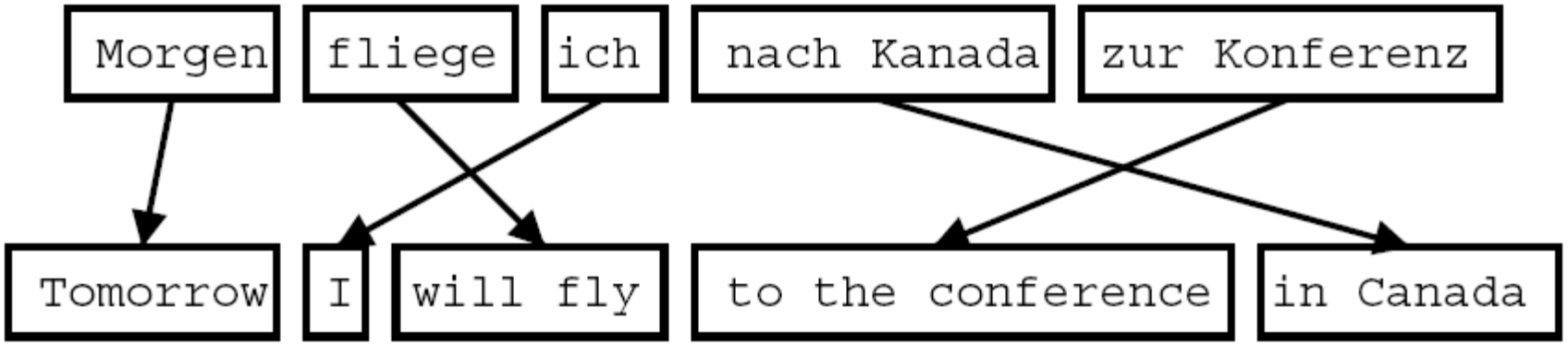
- Spanish–Portuguese
  - Spanish
    - **Todos** los **seres humanos** nacen libres **e** iguales en dignidad y derechos y, **dotados** como están **de** razón y conciencia, deben comportarse fraternalmente los unos con los otros.
  - Portuguese
    - **Todos** os **seres humanos** nascem livres **e** iguais em dignidade e em direitos. **Dotados de** razão e de consciência, devem agir uns para com os outros em espírito de fraternidade.

How Phrase-Based SMT Systems Work

A Very Brief Overview

Phrase-based SMT

1. The sentence is segmented into phrases.
2. Each phrase is translated in isolation.
3. The phrases are reordered.



Phrases: Learnt from a Bi-text

1 Europe's Divided Racial House	1 La Dividida Cámara Racial de
2 A common feature of Europe's e of the immigration issue as a	2 Una característica común de su racismo y su uso del tema política.
3 The Lega Nord in Italy, the Vl supporters of Le Pen's Nationa of parties or movements formed immigrants and promotion of si	3 La Lega Nord de Italia, el V partidarios del Frente Nacio ejemplos de partidos o movim tema común de la aversión a de políticas simplistas para
4 While individuals like Jorg Ha and (never to soon) go, the ra European politics anytime soon	4 Aunque los individuos como J vienen y (nunca demasiado pr raza no desaparecerá de la p momento cercano.
5 An aging population at home ar increasing racial fragmentatic	5 La población cada vez más vi abiertas que nunca, implican los países europeos.
6 Mainstream parties of the cent confronted this prospect by hi hoping against hope that the p	6 Los principales partidos de derecha se han enfrentado a cabeza en la tierra, abrigan problema desaparezca.
7 It will not, as America's raci	7 No lo hará, como claramente
8 Race relations in the US have the center of political debate cleavages are as important as determinants of political pref	
9 The first step to address raci	

Sample Phrase Table

" -- as in ||| " - como en el caso de ||| 1 0.08 0.25 8.04e-07 2.718
" -- as in ||| " - como en el caso ||| 1 0.08 0.25 3.19e-06 2.718
" -- as in ||| " - como en el ||| 1 0.08 0.25 0.003 2.718
" -- as in ||| " - como en ||| 1 0.08 0.25 0.07 2.718

……

is more ||| es mucho más ||| 0.025 0.275 0.007 0.0002 2.718
is more ||| es más un club de ||| 1 0.275 0.007 9.62e-09 2.718
is more ||| es más un club ||| 1 0.275 0.007 3.82e-08 2.718
is more ||| es más un ||| 0.25 0.275 0.007 0.003 2.718
is more ||| es más ||| 0.39 0.275 0.653 0.441 2.718

Using an Additional Related Language

Bi-text Combination Strategies

Problem Definition

- source languages
  - X_1 – resource-poor
  - X_2 - resource-rich
- target language: Y

- We want improved SMT
  - From
    - a resource-poor source language X_1
  - Into
    - a resource-rich target language Y
  - Given
    - a small bi-text for X_1-Y
    - a much larger bi-text for X_2-Y

    for a resource-rich language X_2 closely related to X_1

Bi-text Combination Strategies

- Concatenating bi-texts
- Merging phrase tables
- Our method

Bi-text Combination Strategies

- **Concatenating bi-texts**
- Merging phrase tables
- Our method

Concatenating Bi-texts

- source languages
  - X_1 – resource-poor
  - X_2 - resource-rich
- target language: Y

- Summary: Concatenate X_1-Y and X_2-Y
- Advantages:
  - improved word alignments
    - e.g., for rare words
  - more translation options
    - less unknown words
    - useful non-compositional phrases (improved fluency)
    - phrases with words for language X_2 that do not exist in X_1 are effectively ignored at translation time
- Disadvantages:
  - the additional bi-text X_2-Y will dominate: it is larger
    - translation probabilities are messed up
  - phrases from X_1-Y and X_2-Y cannot be distinguished

Concatenating Bi-texts (2)

- **Concat×*k*:** Concatenate *k* copies of the original and one copy of the additional training bi-text.

- **Concat×*k*:align:**
  1. Concatenate *k* copies of the original and one copy of the additional bi-text.
  2. Generate word alignments.
  3. Truncate them only keeping alignments for one copy of the original bi-text.
  4. Build a phrase table.
  5. Tune the system using MERT.

The value of k is optimized on the development dataset.

Bi-text Combination Strategies

- Concatenating bi-texts
- **Merging phrase tables**
- Our method

Merging Phrase Tables

- source languages
  - X_1 – resource-poor
  - X_2 - resource-rich
- target language: Y

- Summary: Build two separate phrase tables, then
  (a) use them together: as alternative decoding paths
  (b) merge them: using extra features to indicate the bi-text each phrase entry came from
  (c) interpolate them: e.g., using linear interpolation

- Advantages:
  - phrases from X_1-Y and X_2-Y can be distinguished
  - the larger bi-text X_2-Y does not dominate X_1-Y
  - more translation options
  - probabilities are combined in a principled manner

- Disadvantages:
  - improved word alignments are not possible

Merging Phrase Tables (2)

- **Two-tables:** Build two separate phrase tables and use them as alternative decoding paths (Birch et al., 2007).

Merging Phrase Tables (3)

- **Interpolation:** Build two separate phrase tables, T_{orig} and T_{extra}, and combine them using linear interpolation:

  $$\Pr(e|s) = \alpha \Pr_{orig}(e|s) + (1 - \alpha)\Pr_{extra}(e|s).$$

The value of α is optimized on the development dataset, trying the following values: .5, .6, .7, .8, and .9.

Merging Phrase Tables (4)

- **Merge:**
  1. Build separate phrase tables: T_{orig} and T_{extra}.
  2. Keep all entries from T_{orig}.
  3. Add those phrase pairs from T_{extra} that are not in T_{orig}.
  4. Add extra features:
     - F1: 1 if the entry came from T_{orig}, 0.5 otherwise.
     - F2: 1 if the entry came from T_{extra}, 0.5 otherwise.
     - F3: 1 if the entry was in both tables, 0.5 otherwise.

The feature weights are set using MERT, and the number of features is optimized on the development set.

Bi-text Combination Strategies

- Concatenating bi-texts
- Merging phrase tables
- **Our method**

Our Method

Improved word alignments.

- Use **Merge** to combine the phrase tables for **concat×k:align** (as T_{orig}) and for **concat×1** (as T_{extra}).

Distinguish phrases by source table.

Improved lexical coverage.

- Two parameters to tune
  - number of repetitions *k*
  - # of extra features to use with Merge:
    - (a) F1 only;
    - (b) F1 and F2,
    - (c) F1, F2 and F3

Data

Language Pairs

- Use the following language pairs:
  - Indonesian → English (resource-poor)
    - using Malay → English (resource-rich)
  - Spanish → English (resource-poor)
    - using Portuguese → English (resource-rich)

We just pretend that Spanish is resource-poor.

Datasets

- Indonesian-English (in-en):
  - **28,383 sentence pairs** (0.8M, 0.9M words);
  - monolingual English en_{in}: 5.1M words.
- Malay-English (ml-en):
  - **190,503 sentence pairs** (5.4M, 5.8M words);
  - monolingual English en_{ml}: 27.9M words.
- Spanish-English (es-en):
  - **1,240,518 sentence pairs** (35.7M, 34.6M words);
  - monolingual English $en_{es:pt}$: 45.3M words (same for pt-en).
- Portuguese-English (pt-en):
  - **1,230,038 sentence pairs** (35.9M, 34.6M words).
  - monolingual English $en_{es:pt}$: 45.3M words (same for es-en).

Transliteration

Cognate-based
character-level model

Cognates

- Linguistics
  - Def: Words derived from a common root, e.g.,
    - Latin *tu* ('2nd person singular')
    - Old English *thou*
    - French *tu*
    - Spanish *tú*
    - German *du*
    - Greek *sú*

Wordforms can differ:
- *night* vs. *nacht* vs. *nuit* vs. *noite* vs. *noch*
- *star* vs. *estrella* vs. *stella* vs. *étoile*
- *arbeit* vs. *rabota* vs. *robota* ('work')
- *father* vs. *père*
- *head* vs. *chef*

  - Orthography/phonetics/semantics: ignored.

- Computational linguistics
  - Def: Words in different languages that are mutual translations and have a similar orthography, e.g.,
    - *evolution vs. evolución* vs. *evolução*
  - Orthography & semantics: important.
  - Origin: ignored.

Spelling Differences Between Cognates

- **Systematic spelling differences**
  - Spanish – Portuguese
    - different spelling
      - *-nh-* → *-ñ-* (*senhor* vs. *señor*)
    - phonetic
      - *-ción* → *-ção* (*evolución* vs. *evolução*)
      - *-é* → *-ei* (1st sing past) (*visité* vs. *visitei*)
      - *-ó* → *-ou* (3rd sing past) (*visitó* vs. *visitou*)

Many of these differences can be learned automatically.

- **Occasional differences**
  - Spanish – Portuguese
    - *decir* vs. *dizer* ('to say')
    - *Mario* vs. *Mário*
    - *María* vs. *Maria*
  - Malay – Indonesian
    - *kerana* vs. *karena* ('because')
    - *Inggeris* vs. *Inggris* ('English')
    - *mahu* vs. *mau* ('want')

Automatic Transliteration (1)

- Transliteration
  1. Extract likely cognates for Portuguese-Spanish
  2. Learn a character-level transliteration model
  3. Transliterate the Portuguese side of pt-en, to look like Spanish

Automatic Transliteration (2)

- Portuguese-Spanish transliteration using English as a pivot:
  1. Build IBM Model 4 alignments for *pt-en*, *en-es*.
  2. Extract pairs of likely *pt-es* cognates using English (*en*) as a pivot.
  3. Train and tune a character-level SMT system.
  4. Transliterate the Portuguese side of *pt-en*.

Transliteration did not help much for Malay-Indonesian.

Automatic Transliteration (3)

- Extract *pt-es* cognates using English (*en*)
  1. Induce *pt-es* word translation probabilities

$$\Pr(p_j|s_k) = \sum_i \Pr(p_j|e_i)\Pr(e_i|s_k)$$

$$\Pr(s_k|p_j) = \sum_i \Pr(s_k|e_i)\Pr(e_i|p_j)$$

  2. Filter out by probability if

$$\text{Prod}(p_j, s_k) = \Pr(p_j|s_k)\Pr(s_k|p_j) < 0.01$$

  3. Filter out by orthographic similarity if

$$\frac{|\text{LCS}(s_1,s_2)|}{\max(|s_1|,|s_2|)} < 0.58$$

constants proposed in the literature

Longest common subsequence

Automatic Transliteration (4)

- Induce *pt-es* word translation probabilities
  - We can express Pr(p_j|s_k) as follows:

$$\Pr(p_j|s_k) = \sum_i \Pr(p_j|e_i, s_k)\Pr(e_i|s_k)$$

  - Assuming that p_j is conditionally independent of s_k given e_i, we obtain

$$\Pr(p_j|s_k) = \sum_i \Pr(p_j|e_i)\Pr(e_i|s_k)$$

  - This is what was used on the previous slide

Automatic Transliteration (5)

Train & tune a monotone character-level SMT system

- Representation

  ^ e v o l u ç ã o $ — ^ e v o l u c i ó n $

- Data
  - 28,725 pt-es cognate pairs (total)
  - 9,201 (32%) had spelling differences
  - train/tune split: 26,725 / 2,000 pairs
  - language model: 34.6M Spanish words

- Tuning BLEU: 95.22% (baseline: 87.63%)

We use this model to transliterate the Portuguese side of *pt-en*.

Evaluation and Results

Cross-lingual SMT Experiments: Malay & Indonesian

train on Malay, test on Malay

#	Train	Dev	Test	LM	10K	20K	40K	80K	160K
1	ml-en	ml-en	ml-en	en_{ml}	44.93	46.98	47.15	48.04	49.01
2	ml_{in}-en	ml-en	ml-en	en_{ml}	38.99	40.96	41.02	41.88	42.81
3	ml-en	ml-en	**in-en**	en_{ml}	13.69	14.58	14.76	15.12	15.84
4	ml-en	**in-en**	**in-en**	en_{ml}	13.98	14.75	14.91	15.51	16.27
5	ml-en	**in-en**	**in-en**	$\textbf{en}_{in}$	15.56	16.38	16.52	17.04	17.90
6	ml_{in}-en	**in-en**	**in-en**	$\textbf{en}_{in}$	16.44	17.36	17.62	18.14	19.15

train on Malay, test on Indonesian

Cross-lingual SMT Experiments: Spanish & Portuguese

train on Portuguese, test on Portuguese

#	Train	Dev	Test	LM	10K	20K	40K	80K	160K	320K	640K	1.23M
1	pt-en	pt-en	pt-en	$\text{en}_{es:pt}$	21.28	23.11	24.43	25.72	26.43	27.10	27.78	27.96
2	pt_{es}-en	pt-en	pt-en	$\text{en}_{es:pt}$	10.91	11.56	12.16	12.50	12.83	13.27	13.48	13.71
3	pt-en	pt-en	**es-en**	**$\text{en}_{es:pt}$**	4.40	4.77	4.57	5.02	4.99	5.32	5.08	5.34
4	pt-en	**es-en**	**es-en**	**$\text{en}_{es:pt}$**	4.91	5.12	5.64	5.82	6.35	6.87	6.44	7.10
5	pt_{es}-en	**es-en**	**es-en**	**$\text{en}_{es:pt}$**	8.18	9.03	9.97	10.66	11.35	12.26	12.69	13.79

train on Portuguese, test on Spanish

Indonesian→English (using Malay)

Original (constant)

Extra (changing)

Second best method

stat. sign. over baseline

in-en	*ml-en*	Baseline	Two tables	Interpol.	Merge	concat×1	concat×k	concat×k:align	Our method
28.4K	10K	23.80$^{<}$	$^{\geq}$23.79$^{<}$	23.89$^{<}_{(.9)}$	23.97$^{<}_{(3)}$	24.29$^{<}$	24.29$^{<}_{(1)}$	24.01$^{<}_{(1)}$	$^{<}$**24.51**$_{(2;1)}$ (+0.72)
28.4K	20K	23.80$^{<}$	24.24$^{<}$	24.22$^{<}_{(.8)}$	$^{\leq}$24.46$^{<}_{(3)}$	24.37$^{<}$	$^{\leq}$24.48$_{(2)}$	$^{<}$24.35$^{<}_{(2)}$	$^{<}$**24.70**$_{(2;2)}$ (+0.90)
28.4K	40K	23.80$^{<}$	24.27$^{<}$	24.27$^{<}_{(.8)}$	24.43$^{\leq}_{(3)}$	24.38$^{\leq}$	$^{\leq}$24.54$_{(4)}$	$^{<}$24.39$^{<}_{(4)}$	$^{<}$**24.73**$_{(4;2)}$ (+0.93)
28.4K	80K	23.80$^{<}$	24.11$^{<}$	$^{\leq}$24.46$^{<}_{(.8)}$	$^{<}$24.67$_{(3)}$	24.17$^{<}$	$^{\leq}$24.65$^{<}_{(8)}$	24.18$^{<}_{(8)}$	$^{<}$**24.97**$_{(8;3)}$ (+1.17)
28.4K	160K	23.80$^{<}$	$^{<}$24.58$^{<}$	$^{<}$24.58$^{<}_{(.8)}$	$^{<}$24.79$^{\leq}_{(3)}$	$^{\leq}$24.43$^{<}$	$^{<}$25.00$_{(16)}$	$^{\leq}$24.27$^{<}_{(16)}$	$^{<}$**25.15**$_{(16;3)}$ (+1.35)

Spanish→English (using Portuguese)

Orig. (varied)

stat. sign. over baseline

es-en	*pt-en*	Transl.	Baseline	Two tables	Interpol.	Merge	concat×1	concat×k	concat×k:align	Our method
10K	160K	no	$22.87^{<}$	$^{<}23.81$	$^{<}23.73_{(.5)}$	$^{<}23.60_{(2)}$	$^{<}23.54^{<}$	$^{<}23.83^{<}_{(16)}$	$22.93^{<}_{(16)}$	$^{<}$**23.98**$_{(16;3)}$ (+1.11)
		yes	$22.87^{<}$	$^{<}25.29^{\leq}$	$^{<}25.22^{<}_{(.5)}$	$^{<}25.16^{<}_{(2)}$	$^{<}25.26$	$^{<}25.42_{(16)}$	$^{<}23.31^{<}_{(16)}$	$^{<}$**25.73**$_{(16;3)}$ (+2.86)
20K	160K	no	$24.71^{<}$	$^{<}25.22$	$^{\leq}25.02^{<}_{(.5)}$	$^{<}25.32^{\leq}_{(3)}$	$^{<}25.19^{<}$	$^{<}25.29^{<}_{(8)}$	$24.91^{<}_{(8)}$	$^{<}$**25.65**$_{(8;2)}$ (+0.94)
		yes	$24.71^{<}$	$^{<}26.07^{\leq}$	$^{<}26.07_{(.7)}$	$^{<}26.04^{<}_{(3)}$	$^{<}26.16^{\leq}$	$^{<}26.18^{\leq}_{(8)}$	$24.88^{<}_{(8)}$	$^{<}$**26.36**$_{(8;3)}$ (+1.65)
40K	160K	no	$25.80^{<}$	$25.96^{<}$	$26.15^{<}_{(.6)}$	$25.99^{<}_{(3)}$	$26.24^{<}$	$25.92^{<}_{(4)}$	$25.99^{<}_{(4)}$	$^{<}$**26.49**$_{(4;2)}$ (+0.69)
		yes	$25.80^{<}$	$^{<}26.68$	$^{<}26.43_{(.7)}$	$^{<}26.64_{(3)}$	$^{<}26.78$	$^{<}26.93_{(4)}$	$25.88^{<}_{(4)}$	$^{<}$**26.95**$_{(4;3)}$ (+1.15)
80K	160K	no	$27.08^{\leq}$	$^{\geq}26.89^{<}$	$27.04^{<}_{(.8)}$	$27.02^{<}_{(3)}$	27.23	$27.09^{<}_{(2)}$	$27.01^{<}_{(2)}$	$^{\leq}$**27.30**$_{(2;2)}$ (+0.22)
		yes	$27.08^{<}$	$27.20^{<}$	$27.42_{(.5)}$	$27.29^{\leq}_{(3)}$	$27.26^{<}$	$^{\leq}27.53_{(2)}$	$27.09^{<}_{(2)}$	$^{<}$**27.49**$_{(2;3)}$ (+0.41)
160K	160K	no	27.90	27.99	$27.72_{(.5)}$	$27.95_{(2)}$	$27.83^{<}$	$27.83^{<}_{(1)}$	$27.94_{(1)}$	**28.05**$_{(1;3)}$ (+0.15)
		yes	27.90	28.11	$^{\leq}28.13_{(.6)}$	$^{\leq}28.17_{(2)}$	$^{\leq}28.14$	$^{\leq}28.14_{(1)}$	$28.06_{(1)}$	**28.16**$_{(1;2)}$ (+0.26)

Spanish→English (using Portuguese)

System	10K	20K	40K	80K	160K	320K
baseline	22.87	24.71	25.80	27.08	27.90	28.46
our method: **160K** *pt-en* pairs	23.98*	25.65*	26.49*	27.30◇	28.05	28.52
– improvement	**+1.11***	**+0.94***	**+0.69***	**+0.22◇**	**+0.15**	**+0.06**
our method: **1.23M** *pt-en* pairs	24.23*	25.70*	26.78*	27.49	28.22◇	28.58
– improvement	**+1.36***	**+0.99***	**+0.98***	**+0.41**	**+0.32◇**	**+0.12**

Related Work

1. Using Cognates
2. Paraphrasing with a Pivot Language

Using Cognates

- Al-Onaizan et al. (1999) used likely cognates to improve Czech-English word alignments
  (a) by seeding the parameters of IBM model 1
  (b) by constraining word co-occurrences for IBM models 1-4
  (c) by using the cognate pairs as additional "sentence pairs"

- Kondrak et al. (2003): improved SMT for nine European languages
  - using the "sentence pairs" approach

Al-Onaizan & al. vs. Our Method

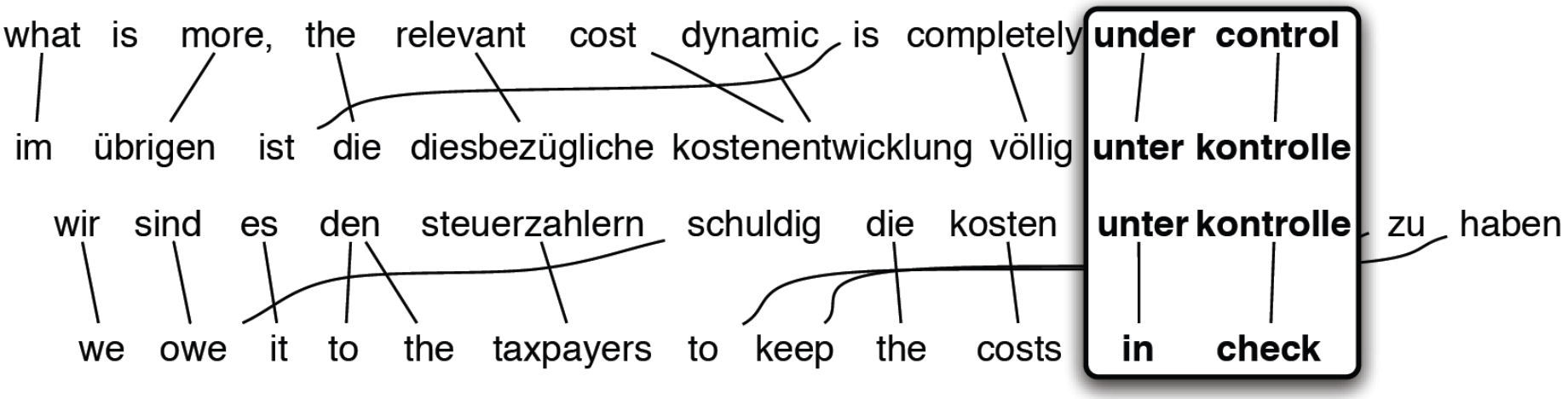
- Use cognates between the source and the target languages
- Extract cognates explicitly
- Do not use context
- Use single words only

- Uses cognates between the source and some non-target language
- Does not extract cognates *(except for transliteration)*
- Leaves cognates in their sentence contexts
- Can use multi-word cognate phrases

The two approaches are orthogonal and thus can be combined.

Paraphrasing Using a Pivot Language

- *Paraphrasing with Bilingual Parallel Corpora* (Bannard & Callison-Burch'05)
- *Improved statistical machine translation using paraphrases* (Callison-Burch &al.'06)



e.g., Spanish→English (using German)

Improved MT Using a Pivot Language

- Use many pivot languages
- New source phrases are added to the phrase table (paired with the original English)
- A new feature is added to each table entry:

$$h(\mathbf{e}, \mathbf{f_1}) = \begin{cases} p(\mathbf{f_2}|\mathbf{f_1}) & \text{If phrase table entry } (\mathbf{e}, \mathbf{f_1}) \\ & \text{is generated from } (\mathbf{e}, \mathbf{f_2}) \\ 1 & \text{Otherwise} \end{cases}$$

Pivoting vs. Our Approach

– can only improve source-language lexical coverage	+ augments both the source- and the target-language sides
– ignores context entirely	+ takes context into account
+ the additional language does not have to be related to the source	– requires that the additional language be related to the source

The two approaches are orthogonal and thus can be combined.

Conclusion and Future Work

Overall

- We have presented
  - An approach that uses a bi-text for a resource-rich language pair to build a better SMT system for a related resource-poor language.
- We have achieved
  - Up to 3.37 Bleu points absolute improvement for Spanish-English (using Portuguese)
  - Up to 1.35 for Indonesian-English (using Malay)
- The approach could be used for many resource-poor languages.

Future Work

- Try auxiliary languages related to the *target.*
- Extend the approach to a multi-lingual corpus, e.g., Spanish-Portuguese-English.

Thank You

Any questions?

This research was supported
by research grant POD0713875.