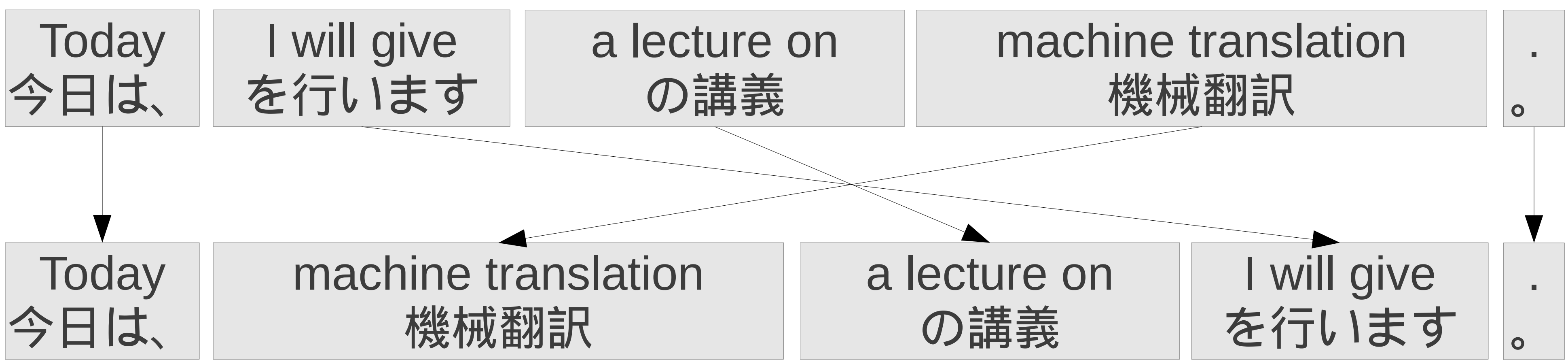
Building a Phrase-based SMT System

Graham Neubig & Kevin Duh
Nara Institute of Science and Technology (NAIST)
5/10/2012

Phrase-based
Statistical Machine Translation (SMT)

- Divide sentence into patterns, reorder, combine

Today I will give a lecture on machine translation .



今日は、機械翻訳の講義を行います。

- Statistical translation models, reordering models, language models learned from text

This Talk

1) What are the steps required to build a phrase-based machine translation translation system?

2) What tools implement these steps in Moses* (an open-source statistical MT system)?

3) What are some research problems related to each of these components?

* http://www.statmt.org/moses

Steps in Training a Phrase-based SMT System

- Collecting Data
- Tokenization
- Language Modeling
- Alignment
- Phrase Extraction/Scoring
- Reordering Models
- Decoding
- Evaluation
- Tuning

Collecting Data

Collecting Data

- Sentence parallel data
  - Used in: Translation model/Reordering model

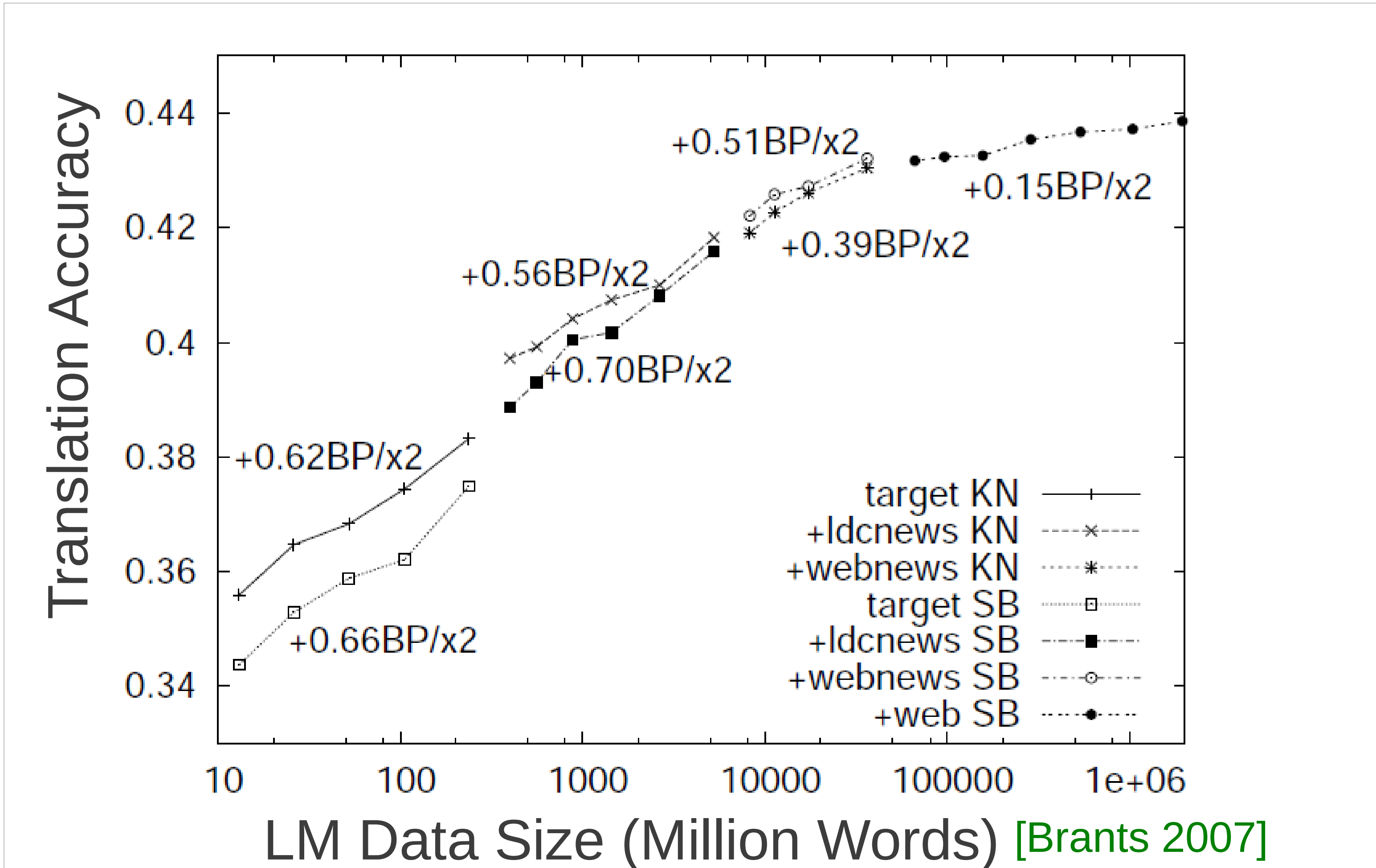
これはペンです。	This is a pen.
昨日は友達と食べた。	I ate with my friend yesterday.
象は花が長い。	Elephants' trunks are long.

- Monolingual data (in the target language)
  - Used in: Language model

This is a pen.
I ate with my friend yesterday.
Elephants' trunks are long.

Good Data is

- Big! →



[Brants 2007]

- Clean
- In the same domain as test data

Collecting Data

- For academic workshops, data is prepared for us!

e.g.
IWSLT 2011 →

Name	Type	Words
TED	Lectures	1.76M
News Commentary	News	2.52M
EuroParl	Political	45.7M
UN	Political	301M
Giga	Web	576M

- In real systems
  - Data from government organizations, newspapers
  - Crawl the web
  - Merge several data sources

Research

- Finding bilingual pages [Resnik 03]

毎日jp ホーム ニュース オピニオン スポーツ エンタメ 地域 特集·連載 ENG

オピニオン 社説 余録 解説 コラム

トップ > オピニオン > 記事

[PR] 休肝日が気になる40代男性が始めた健康法！しじみ習慣／無料サンプル

+1 0 ツイート 23 おすすめ 15 チェック 記事を印刷 文字サ

社説:超高齢社会　「肩車型」の常識を疑え

毎日新聞　2012年05月05日　02時30分

長寿はおめでたいことなのに、高齢化となると悲観論をもって語られることが多い。現役世代

続けているせいでもある。現役4人が高齢者1人を背負う「騎馬戦型」から、現役1人が高齢者1

車型」になると言われたら誰しも不安になるだろう。たしかに人口比率はそのようになる。

だからこそ先進国最低レベルの国民負担率（税と保険の負担）をもう少し引き上げるべきだ

「肩車型」説は登場したはずだったが、野田佳彦首相らの言い方がまずいのだろうか、逆に社会

The Mainichi

[PR] ４０歳からの「しじみ習慣」休肝日が気になるあなたに！／無料サンプル

+1 0 ツイート 0 おすすめ チェック 記事を印刷 文字サイズ 小 中 大

Editorial: Aging society does not necessarily spell doom

Longevity is something to be celebrated, but when it comes to the aging of Japanese society, it is often discussed in a pessimistic tone.

One reason for this is the continuing decline in people of working age. Learning that our society is shifting from one in which four working people financially support one senior citizen, to another in which each working person must support one senior citizen -- a so-called "piggyback" setup -- would make anyone anxious. And indeed, that is exactly what is happening.

[Image: Mainichi Shimbun]

Research

- Finding bilingual pages [Resnik 03]
- Sentence alignment [Moore 02]


毎日jp
ホーム ニュース オピニオン スポーツ エンタメ 地域 特集·連載 ENG
オピニオン 社説 余録 解説 コラム
トップ > オピニオン > 記事
[PR] 休肝日が気になる40代男性が始めた健康法！しじみ習慣／無料サンプル
+1 0 ツイート 23 おすすめ 15 チェック 記事を印刷 文字サ

社説:超高齢社会　「肩車型」の常識を疑え
毎日新聞　2012年05月05日　02時30分

長寿はおめでたいことなのに、高齢化となると悲観論をもって語られることが多い。現役世代
続けているせいでもある。現役4人が高齢者1人を背負う「騎馬戦型」から、現役1人が高齢者1
車型」になると言われたら誰しも不安になるだろう。たしかに人口比率はそのようになる。

だからこそ先進国最低レベルの国民負担率（税と保険の負担）をもう少し引き上げるべきだ
「肩車型」説は登場したはずだったが、野田佳彦首相らの言い方がまずいのだろうか、逆に社会

The Mainichi
[PR] ４０歳からの「しじみ習慣」休肝日が気になるあなたに!／無料サンプル
+1 0 ツイート 0 おすすめ チェック 記事を印刷 文字サイズ 小 中 大

Editorial: Aging society does not necessarily spell doom

Longevity is something to be celebrated, but when it comes to the aging of Japanese society, it is often discussed in a pessimistic tone.

One reason for this is the continuing decline in people of working age. Learning that our society is shifting from one in which four working people financially support one senior citizen, to another in which each working person must support one senior citizen -- a so-called "piggyback" setup -- would make anyone anxious. And indeed, that is exactly what is happening.

Research

- Finding bilingual pages [Resnik 03]
- Sentence alignment [Moore 02]


毎日jp ホーム ニュース オピニオン スポーツ エンタメ 地域 特集・連載 ENG

オピニオン 社説 余録 解説 コラム

トップ > オピニオン > 記事

社説:超高齢社会 「肩車型」の常識を疑え

毎日新聞 2012年05月05日 02時30分

長寿はおめでたいことなのに、高齢化となると悲観論をもって語られることが多い。現役世代
続けているせいでもある。現役4人が高齢者1人を背負う「騎馬戦型」から、現役1人が高齢者1
車型」になると言われたら誰しも不安になるだろう。たしかに人口比率はそのようになる。

だからこそ先進国最低レベルの国民負担率(税と保険の負担)をもう少し引き上げるべきだ
「肩車型」説は登場したはずだったが、野田佳彦首相らの言い方がまずいのだろうか、逆に社

The Mainichi

Editorial: Aging society does not necessarily spell doom

Longevity is something to be celebrated, but when it comes to the aging of Japanese society, it is often discussed in a pessimistic tone.

One reason for this is the continuing decline in people of working age. Learning that our society is shifting from one in which four working people financially support one senior citizen, to another in which each working person must support one senior citizen -- a so-called "piggyback" setup -- would make anyone anxious. And indeed, that is exactly what is happening.

- Crowd-sourcing data creation [Ambati 10]
  - Mechanical Turk, duolingo, etc.

Tokenization

Tokenization

- Example: Divide Japanese into words

太郎が花子を訪問した。

↓

太郎 が 花子 を 訪問 した 。

- Example: Make English lowercase, split punctuation

Taro visited Hanako.

↓

taro visited hanako .

Tools for Tokenization

- Most European languages

```
tokenize.perl en < input.en > output.en
tokenize.perl fr < input.fr > output.fr
```

- Japanese

MeCab: `mecab -O wakati < input.ja > output.ja`
KyTea: `kytea -notags < input.ja > output.ja`
JUMAN, etc.

- Chinese

Stanford Segmenter, LDC, KyTea, etc...

Research

- What is good tokenization for machine translation?
  - Accuracy? Consistency? [Chang 08]
  - Matching target language words? [Sudoh 11]

太郎 が 花子 を 訪問 した 。

Taro <ARG1> visited <ARG2> Hanako .

- Morphology (Korean, Arabic, Russian) [Niessen 01]

단어란 도대체 무엇일까요?

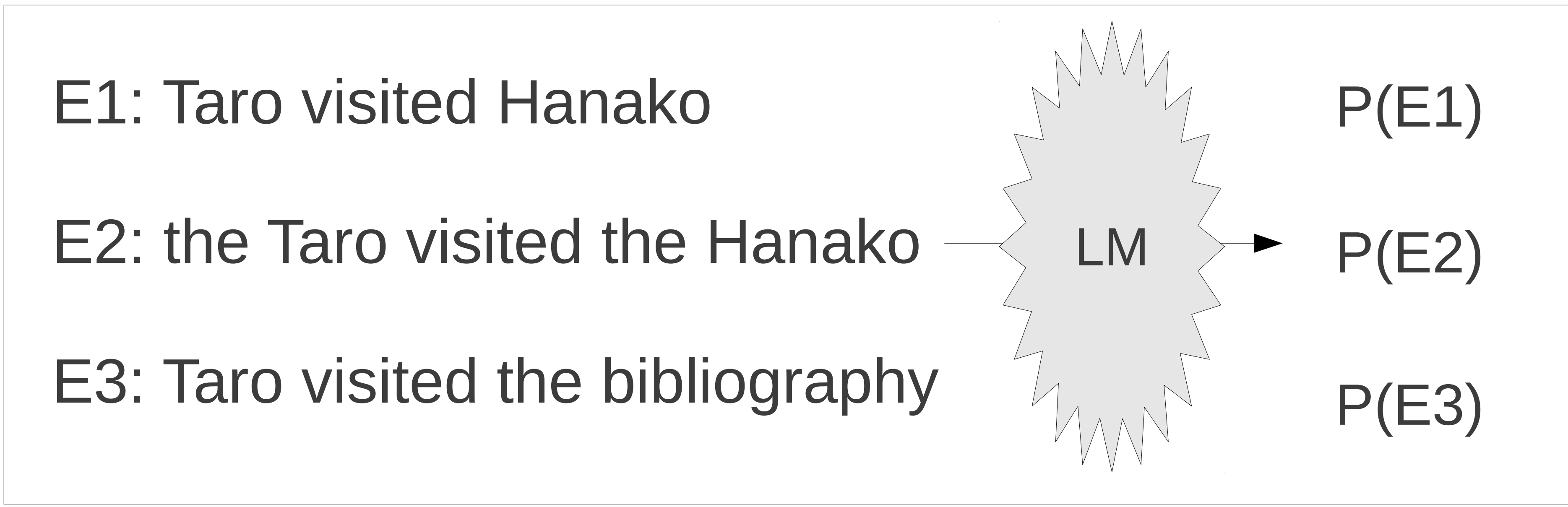
단어 란 도대체 무엇 일 까요 ?

- Unsupervised learning [Chung 09, Neubig 12]

Language Modeling

Language Modeling

- Assign a probability to each sentence



- More fluent sentences get higher probability

P(E1) > P(E2) P(E1) > P(E3)

n-gram Models

- We want the probability of

P(W = “Taro visited Hanako”)

- n-gram model calculates one word at a time
  - Condition on n-1 previous words
    e.g. 2-gram model

$P(w_1=\text{“Taro”}) * P(w_2=\text{”visited”} \mid w_1=\text{“Taro”})$

$* P(w_3=\text{”Hanako”} \mid w_2=\text{”visited”})$

$* P(w_4=\text{”</s>”} \mid w_3=\text{”Hanako”})$

NOTE:
sentence ending symbol </s>

Tools

- SRILM Toolkit:

  Train:

```
ngram-count -order 5 -interpolate -kndiscount -unk
     -text input.txt -lm lm.arpa
```

  Test:

```
ngram -lm lm.arpa -ppl test.txt
```

- Others: KenLM, RandLM, IRSTLM

Research Problems

- Is there anything that can beat n-grams? [Goodman 01]
  - Fast to compute
  - Easy to integrate into decoding
  - Surprisingly strong
- Other methods
  - Syntactic LMs [Charniak 03]
  - Neural networks [Bengio 06]
  - Model M [Chen 09]
  - etc...

Alignment

Alignment

- Find which words correspond to each-other

太郎 が 花子 を 訪問 した 。

taro visited hanako .

→

太郎 が 花子 を 訪問 した 。

taro visited hanako .

- Done automatically with probabilistic methods

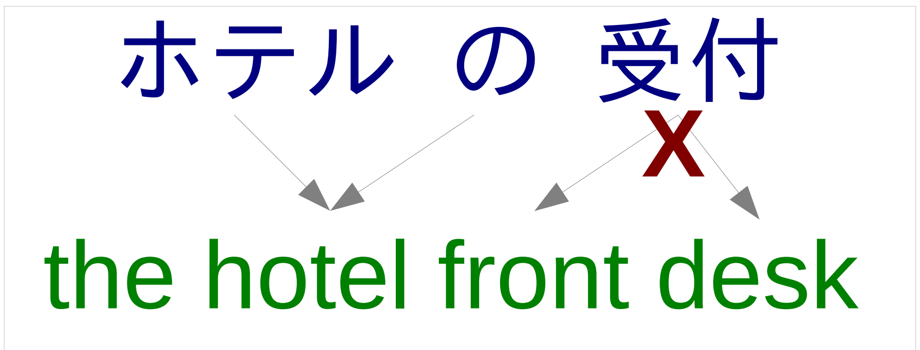
日本語
日本語
日本語

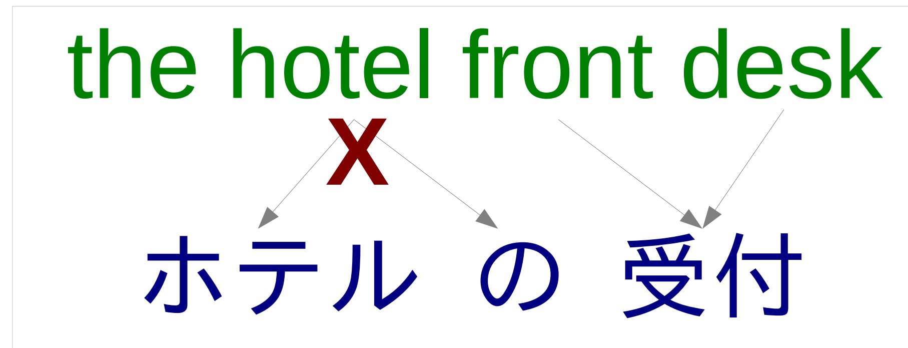
English
English
English

P(花子 |hanako) = 0.99
P(太郎 |taro) = 0.97
P(visited| 訪問) = 0.46
P(visited| した) = 0.04
P(花子 |taro) = 0.0001

IBM/HMM Models

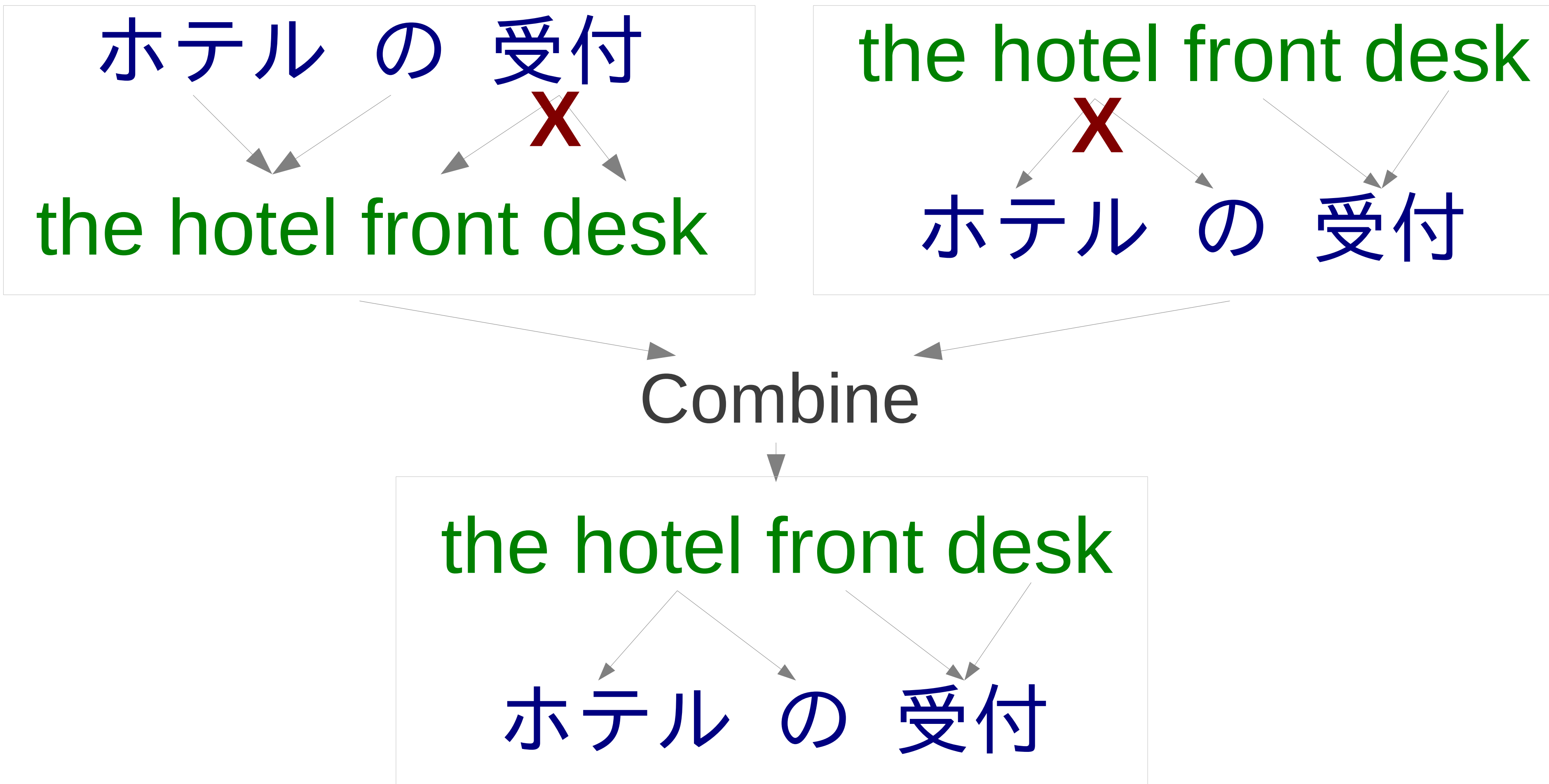
- One-to-many alignment model





  - IBM Model 1: No structure (“bag of words”)
  - IBM Models 2-5, HMM: Add more structure

Combining One-to-Many Alignments



- Several different heuristics

Tools

- mkcls: Find bilingual classes

ホテル の 受付 → 35 49 12

the hotel front desk → 23 35 12 19

- GIZA++: Find alignments using IBM models (uses classes from mkcls for smoothing)

ホテル の 受付 + 35 49 12 → ホテル の 受付

the hotel front desk + 23 35 12 19 → the hotel front desk

- symal: Combine alignments in both directions
- (Included in train-model.perl of Moses)

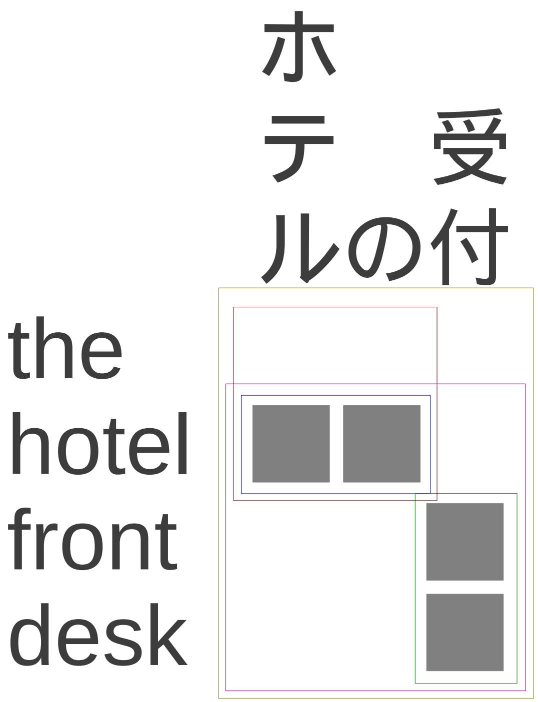
Research Problems

- Does alignment actually matter? [Aryan 06]
- Supervised alignment models [Fraser 06, Haghighi 09]
- Alignment using syntactic structure [DeNero 07]
- Phrase-based alignment models [Marcu 02, DeNero 08]

Phrase Extraction

Phrase Extraction

- Use alignments to find phrase pairs



ホテル の → hotel
ホテル の → the hotel
受付 → front desk
ホテルの受付 → hotel front desk
ホテルの受付 → the hotel front desk

Phrase Scoring

- Calculate 5 standard features
  - Phrase Translation Probabilities:
    P(**f**|**e**) = c(**f**,**e**)/c(**e**)   P(**e**|**f**) = c(**f**,**e**)/c(**f**)

    e.g. c(ホテル の , the hotel) / c(the hotel)
  - Lexical Translation Probabilities
    - Use word-based translation probabilities (IBM Model 1)
    - Helps with sparsity

    $P(f|e) = \Pi_f 1/|\mathbf{e}| \sum_e P(f|e)$

    e.g.
    (P(ホテル |the)+P(ホテル |hotel))/2 * (P(の |the)+P(の |hotel))/2
  - Phrase penalty: 1 for each phrase

Tools

- extract: Extract all the phrases
- phrase-extract/score: Score the phrases
- (Included in train-model.perl)

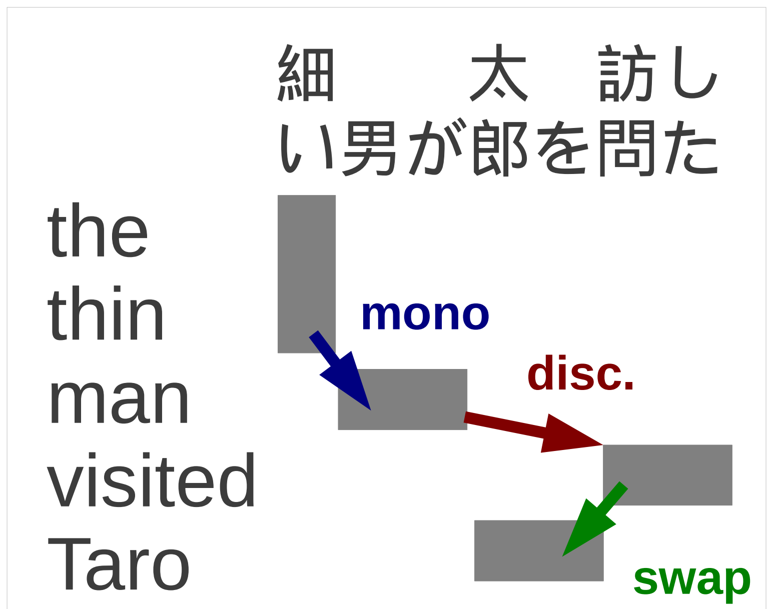
Research

- Domain adaptation of translation models [Koehn 07, Matsoukas 09]
- Reducing phrase table size [Johnson 07]
- Generalized phrase extraction (Geppetto toolkit) [Ling 10]
- Phrase sense disambiguation [Carpuat 07]

Reordering Models

Lexicalized Reordering

- Probability of monotone, swap, discontinuous



細い → the thin
high monotone probability

太郎 を → Taro
high swap probability

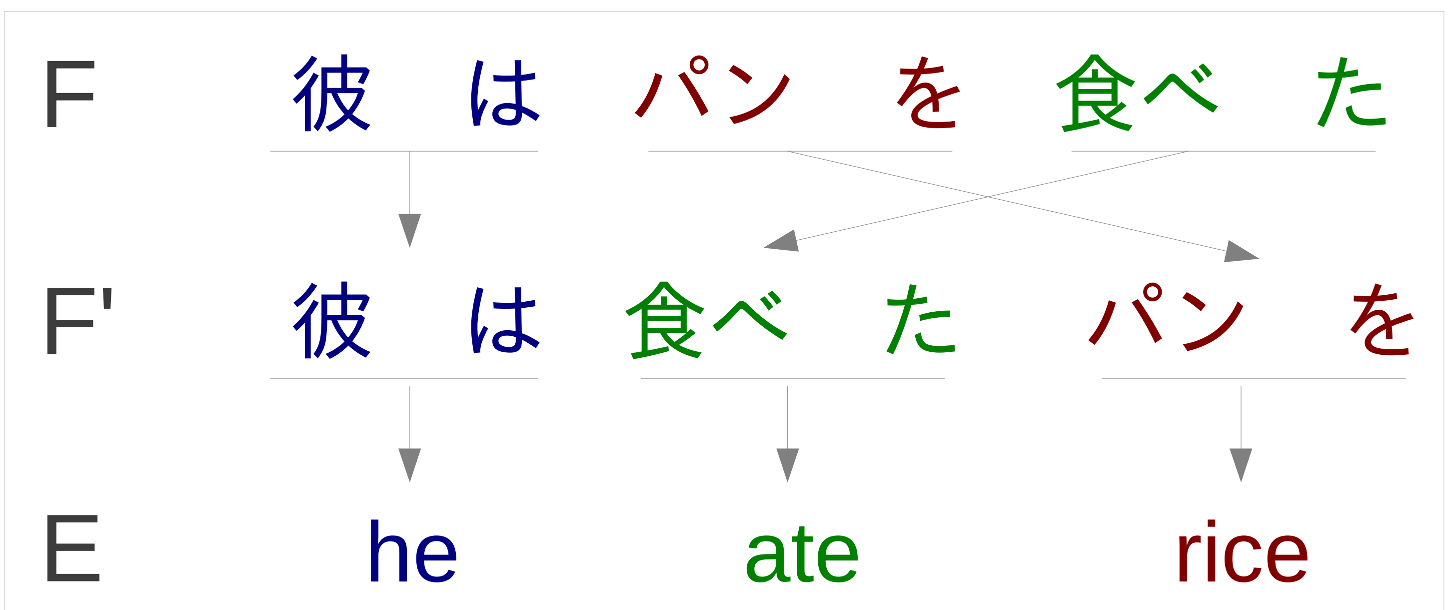
- Conditioning on input/output, left/right, or both

Tools

- extract: Same as phrase extraction
- lexical-reordering/score: Scores lexical reordering
- (included in train-model.perl)

Research

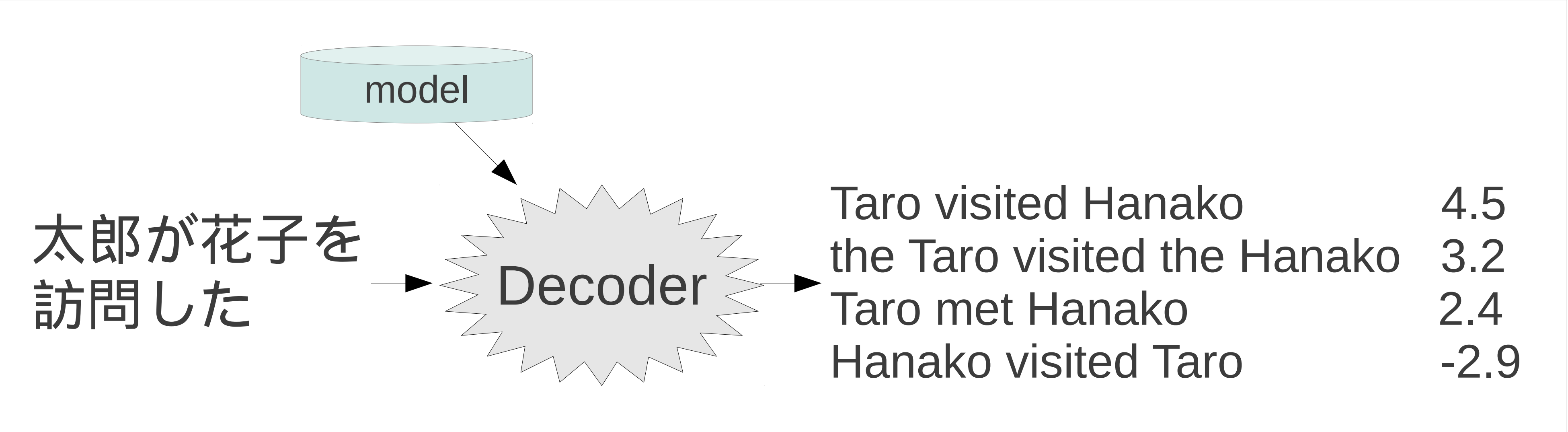
- Still a very open research area (especially en↔ja)
- Change the translation model
  - Hierarchical phrase-based [Chiang 07]
  - Syntax-based translation [Yamada 01, Galley 06]
- Pre-ordering [Xia 04, Isozaki 10]



Decoding

Decoding

- Given the models, find the best answer (or n-best)



- Exact search is NP-hard! [Knight 99]
- Decoding uses beam-search to find an approximate solution [Koehn 03]

Tools

- Moses!

```
moses -f moses.ini < input.txt > output.txt
```

- Also: moses_chart, cdec (for Hiero, syntax-based models)

Research

- Decoding for lattice input [Dyer 08]
- Decoding for syntax models [Mi 08]
- Minimum Bayes risk decoding [Kumar 04]
- Exact decoding [Germann 01]

Evaluation

Human Evaluation

- Adequacy: Is the meaning correct?
- Fluency: Is the sentence natural?
- Pairwise: Is X a better translation than Y?

太郎が花子を訪問した

	Taro visited Hanako	the Taro visited the Hanako	Hanako visited Taro
Adequate?	○	○	X
Fluent?	○	X	○
Better?	B, C	C	

Automatic Evaluation

- How well does the translation match a reference?
  - (or multiple references: more than one correct translation)
- BLEU: n-gram precision, brevity penalty [Papineni 03]

Reference: Taro visited Hanako

System: the Taro visited the Hanako

1-gram: 3/5
2-gram: 1/4
brevity penalty = 1.0

Brevity: min(1, |System|/|Reference|) = min(1, 5/3)

BLEU-2 $= (3/5*1/4)^{1/2} * 1.0$
$= 0.387$

- Also METEOR (normalizes synonyms), TER (# of changes), RIBES (reordering)

Research

- Metrics with focus on a particular thing
  - Reordering [Isozaki 10]
  - Accuracy of meaning [Lo 11]
- Tunable metrics [Cer 10]
- Metric aggregation [Albrecht 07]
- Crowdsourcing human evaluation [Callison-Burch 11]

Tuning

Tuning

- Scores of translation, reordering, and language models

	LM	TM	RM	
○ Taro visited Hanako	-4	-3	-1	-8
X the Taro visited the Hanako	-5	-4	-1	-10
X Hanako visited Taro	-2	-3	-2	-7

Best Score X

- If we add weights, we can get better answers:

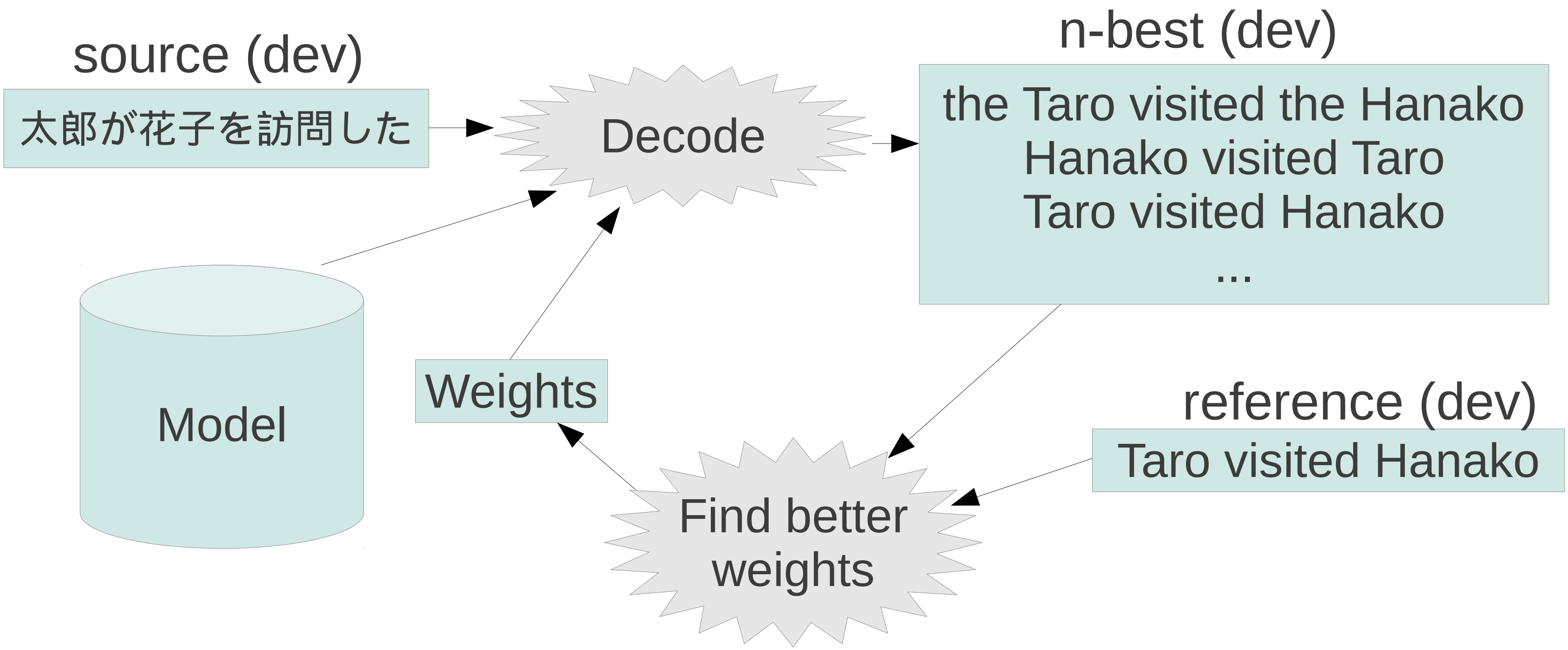
Best Score ○

	LM	TM	RM	
○ Taro visited Hanako	0.2*-4	0.3*-3	0.5*-1	-2.2
X the Taro visited the Hanako	0.2*-5	0.3*-4	0.5*-1	-2.7
X Hanako visited Taro	0.2*-2	0.3*-3	0.5*-2	-2.3

- Tuning finds these weights: w_{LM}=0.2 w_{TM}=0.3 w_{RM}=0.5

Tuning Methods

- Minimum error rate training: MERT [Och 03]



- Others: MIRA [Watanabe 07] (online update), PRO (ranking) [Hopkins 11]

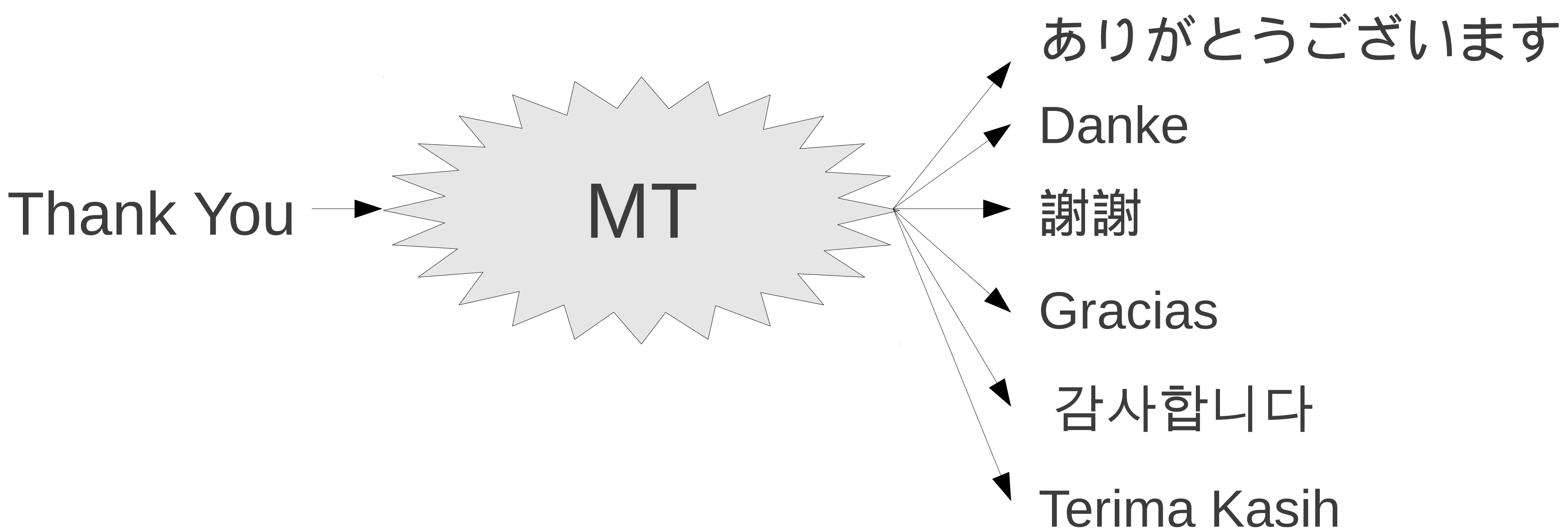
Research

- Tuning with millions of features (e.g. MIRA, PRO)
- Tuning with lattices [Macherey 08]
- Speeding up tuning [Suzuki 11]
- Tuning with multiple metrics [Duh 12]

Last Words

Last Words

- MT is fun! Join us.
- Improving very quickly, but still many problems.
- System is big, but you can focus on one problem.



Bibliography

- J. Albrecht and R. Hwa. A re-examination of machine learning approaches for sentence-level mt evaluation. In Proc. ACL, pages 880-887, 2007.
- V. Ambati, S. Vogel, and J. Carbonell. Active learning and crowdsourcing for machine translation. Proc. LREC, 7:2169-2174, 2010.
- N. Ayan and B. Dorr. Going beyond AER: an extensive analysis of word alignments and their impact on MT. In Proc. ACL, 2006.
- Y. Bengio, H. Schwenk, J.-S. Sencal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. In Innovations in Machine Learning, volume 194, pages 137-186. 2006.
- T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean. Large language models in machine translation. In Proc. EMNLP, pages 858-867, 2007.
- C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan. Findings of the 2011 workshop on statistical machine translation. In Proc. WMT, pages 22-64, 2011.
- M. Carpuat and D. Wu. How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. In Proc. TMI, pages 43-52, 2007.
- D. Cer, C. Manning, and D. Jurafsky. The best lexical metric for phrasebased statistical MT system optimization. In NAACL HLT, 2010.
- P.-C. Chang, M. Galley, and C. D. Manning. Optimizing Chinese word segmentation for machine translation performance. In Proc. WMT, 2008.
- E. Charniak, K. Knight, and K. Yamada. Syntax-based language models for statistical machine translation. In MT Summit IX, pages 40-46, 2003.
- S. Chen. Shrinking exponential language models. In Proc. NAACL, pages 468-476, 2009.
- D. Chiang. Hierarchical phrase-based translation. Computational Linguistics, 33(2), 2007.
- T. Chung and D. Gildea. Unsupervised tokenization for machine translation. In Proc. EMNLP, 2009.
- J. DeNero, A. Bouchard-C^ote, and D. Klein. Sampling alignment structure under a Bayesian translation model. In Proc. EMNLP, 2008.
- J. DeNero and D. Klein. Tailoring word alignments to syntactic machine translation. In Proc. ACL, volume 45, 2007.
- K. Duh, K. Sudoh, X. Wu, H. Tsukada, and M. Nagata. Learning to translate with multiple objectives. In Proc. ACL, 2012.
- C. Dyer, S. Muresan, and P. Resnik. Generalizing word lattice translation. In Proc. ACL, 2008.

- A. Fraser and D. Marcu. Semi-supervised training for statistical word alignment. In Proc. ACL, pages 769-776, 2006.
- M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer. Scalable inference and training of context-rich syntactic translation models. In Proc. ACL, pages 961-968, 2006.
- U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada. Fast decoding and optimal decoding for machine translation. In Proc. ACL, pages 228-235, 2001.
- J. T. Goodman. A bit of progress in language modeling. Computer Speech & Language, 15(4), 2001.
- A. Haghighi, J. Blitzer, J. DeNero, and D. Klein. Better word alignments with supervised ITG models. In Proc. ACL, 2009.
- M. Hopkins and J. May. Tuning as ranking. In Proc. EMNLP, 2011.
- H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada. Automatic evaluation of translation quality for distant language pairs. In Proc. EMNLP, pages 944-952, 2010.
- H. Isozaki, K. Sudoh, H. Tsukada, and K. Duh. Head nalization: A simple reordering rule for sov languages. In Proc. WMT and MetricsMATR, 2010.
- J. H. Johnson, J. Martin, G. Foster, and R. Kuhn. Improving translation quality by discarding most of the phrasetable. In Proc. EMNLP, pages 967-975, 2007.
- K. Knight. Decoding complexity in word-replacement translation models. Computational Linguistics, 25(4), 1999.
- P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In Proc. HLT, pages 48-54, 2003.
- P. Koehn and J. Schroeder. Experiments in domain adaptation for statistical machine translation. In Proc. WMT, 2007.
- S. Kumar and W. Byrne. Minimum bayes-risk decoding for statistical machine translation. In Proc. HLT, 2004.
- W. Ling, T. Lus, J. Graca, L. Coheur, and I. Trancoso. Towards a General and Extensible Phrase-Extraction Algorithm. In M. Federico, I. Lane, M. Paul, and F. Yvon, editors, Proc. IWSLT, pages 313-320, 2010.
- C.-k. Lo and D. Wu. Meant: An inexpensive, high-accuracy, semiautomatic metric for evaluating translation utility based on semantic roles. In Proc. ACL, pages 220-229, 2011.
- W. Macherey, F. Och, I. Thayer, and J. Uszkoreit. Lattice-based minimum error rate training for statistical machine translation. In Proc. EMNLP, 2008.
- D. Marcu and W. Wong. A phrase-based, joint probability model for statistical machine translation. In Proc. EMNLP, 2002.

- S. Matsoukas, A.-V. I. Rosti, and B. Zhang. Discriminative corpus weight estimation for machine translation. In Proc. EMNLP, pages 708717, 2009.
- H. Mi, L. Huang, and Q. Liu. Forest-based translation. In Proc. ACL, pages 192-199, 2008.
- R. Moore. Fast and accurate sentence alignment of bilingual corpora. Machine Translation: From Research to Real Users, pages 135-144, 2002.
- G. Neubig, T. Watanabe, S. Mori, and T. Kawahara. Machine translation without words through substring alignment. In Proc. ACL, Jeju, Korea, 2012.
- S. Niessen, H. Ney, et al. Morpho-syntactic analysis for reordering in statistical machine translation. In Proc. MT Summit, 2001.
- F. J. Och. Minimum error rate training in statistical machine translation. In Proc. ACL, 2003.
- K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In Proc. COLING, pages 311-318, 2002.
- P. Resnik and N. A. Smith. The web as a parallel corpus. Computational Linguistics, 29(3):349-380, 2003.
- J. Suzuki, K. Duh, and M. Nagata. Distributed minimum error rate training of smt using particle swarm optimization. In Proc. IJCNLP, pages 649-657, 2011.
- T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki. Online largemargin training for statistical machine translation. In Proc. EMNLP, pages 764-773, 2007.
- F. Xia and M. McCord. Improving a statistical MT system with automatically learned rewrite patterns. In Proc. COLING, 2004.
- K. Yamada and K. Knight. A syntax-based statistical translation model. In Proc. ACL, 2001.
- O. F. Zaidan and C. Callison-Burch. Crowdsourcing translation: Professional quality from non-professionals. In Proc. ACL, pages 1220-1229, 2011.