TTIC 31190: Natural Language Processing

Kevin Gimpel

Winter 2016

Lecture 16: Machine Translation and other NLP Applications

Announcements

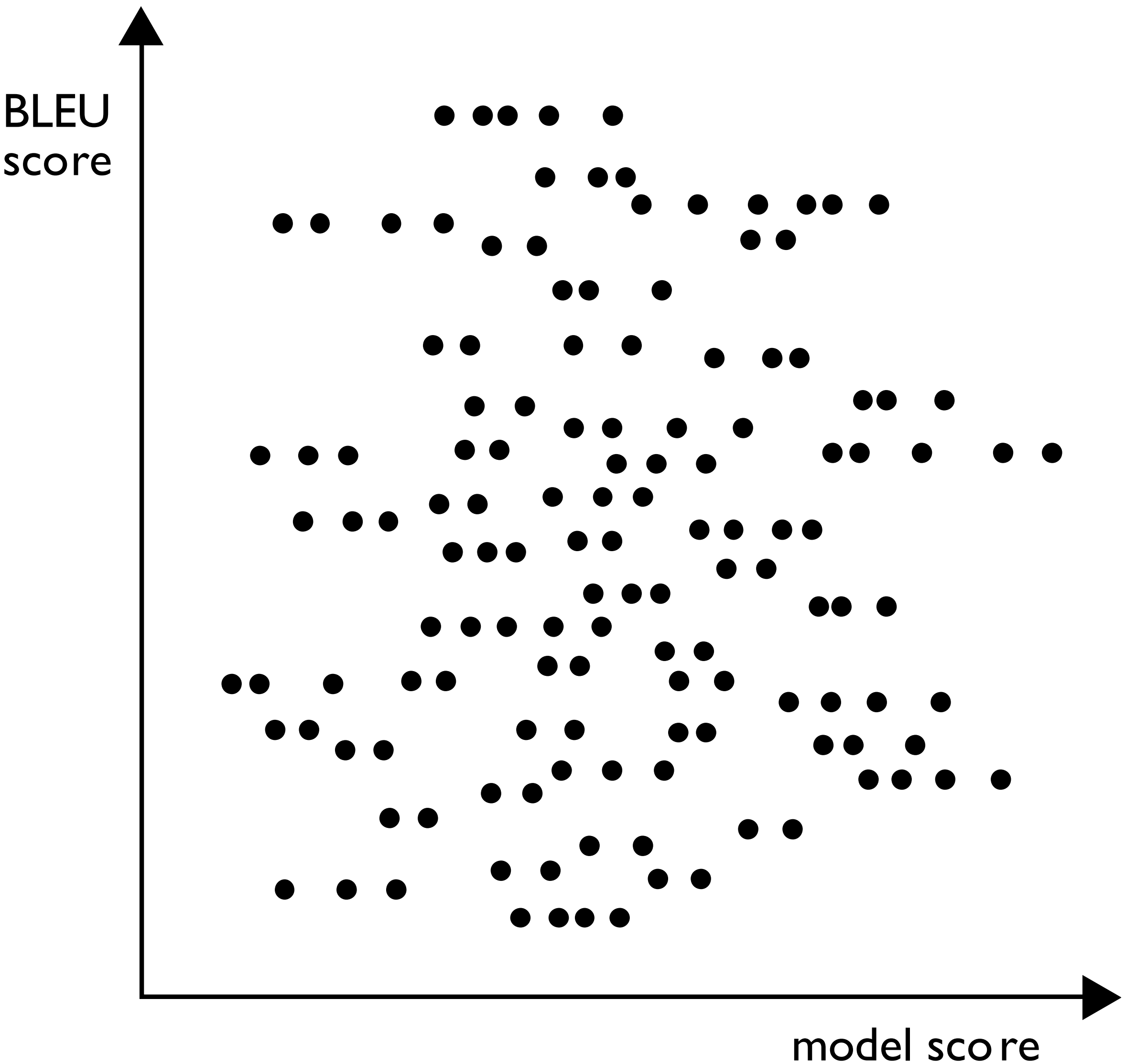
- presentations will actually be 9 minutes because we have so many to fit in
- I will post guidelines on the final project report – think of it as a short (4-page) paper
- I will send you your midterm and assignment 2 grades tomorrow

Roadmap

- classification
- words
- lexical semantics
- language modeling
- sequence labeling
- neural network methods in NLP
- syntax and syntactic parsing
- computational semantics
- machine translation
- other NLP applications

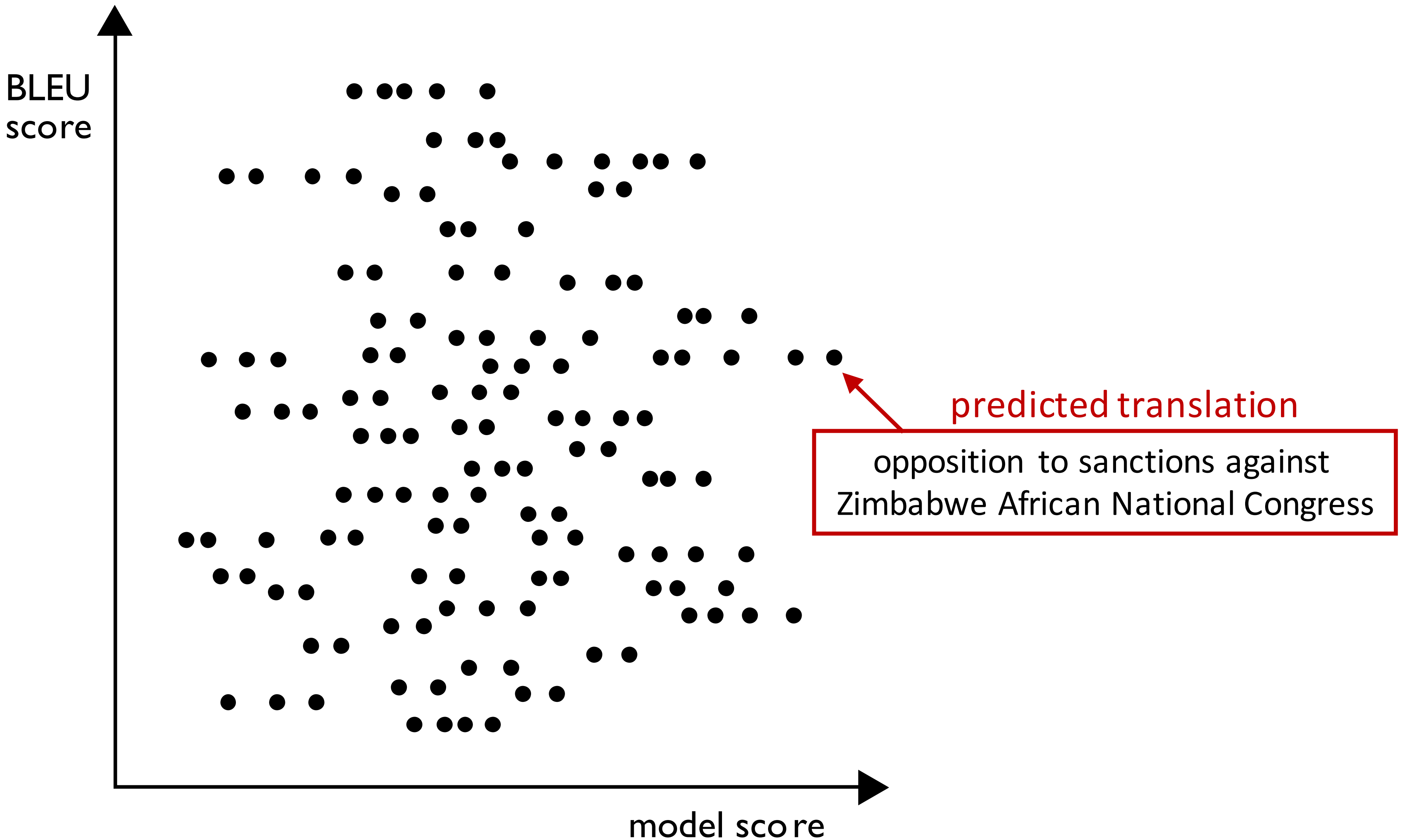
African National Congress 非国大 | opposition 反对 | sanction 制裁 | Zimbabwe 津巴布韦

Gold standard:
African National Congress opposes sanctions against Zimbabwe



African National Congress | opposition | sanction | Zimbabwe
非国大 反对 制裁 津巴布韦

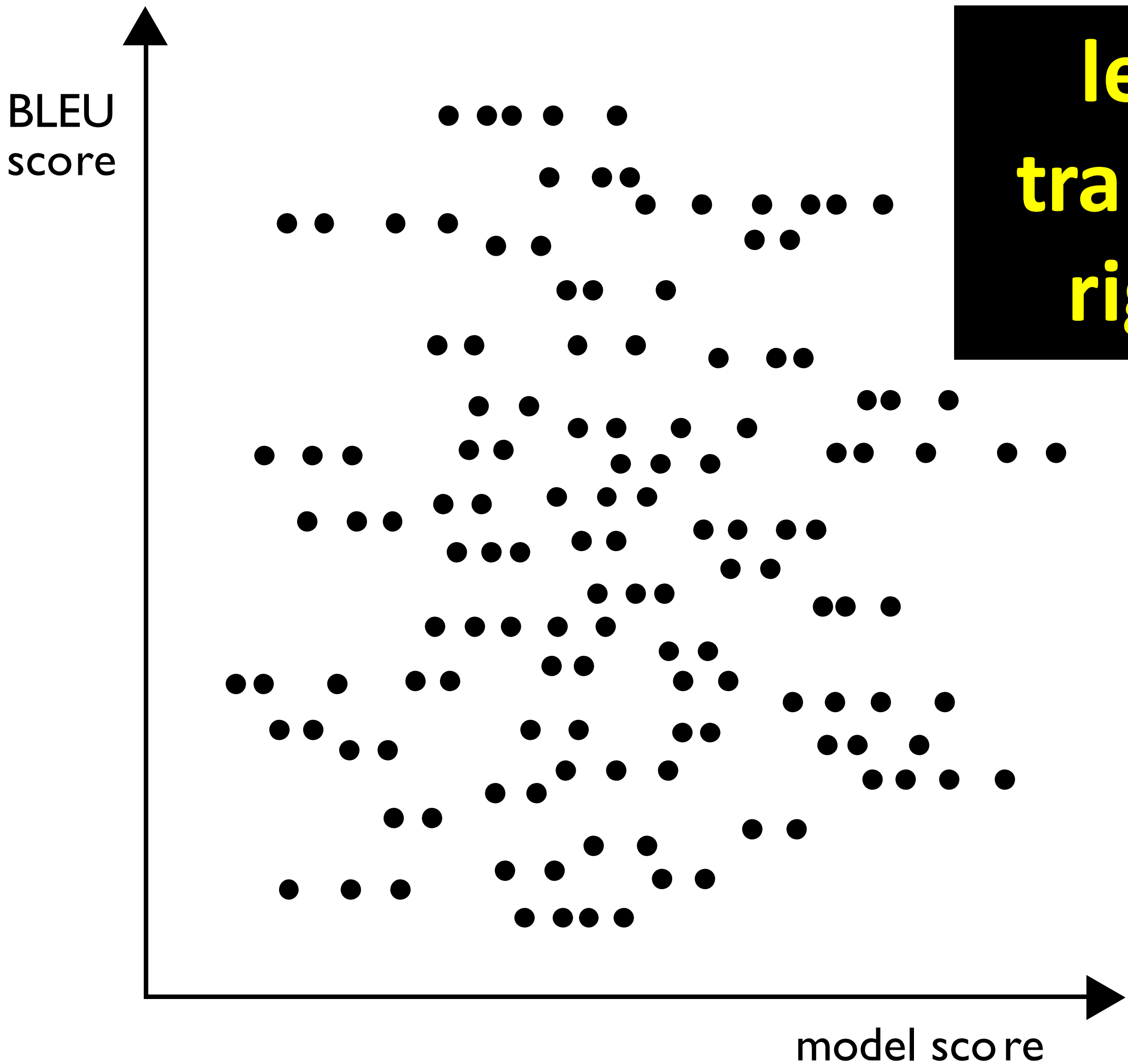
Gold standard:
African National Congress opposes sanctions against Zimbabwe



African National Congress | opposition | sanction | Zimbabwe

非国大 反对 制裁 津巴布韦

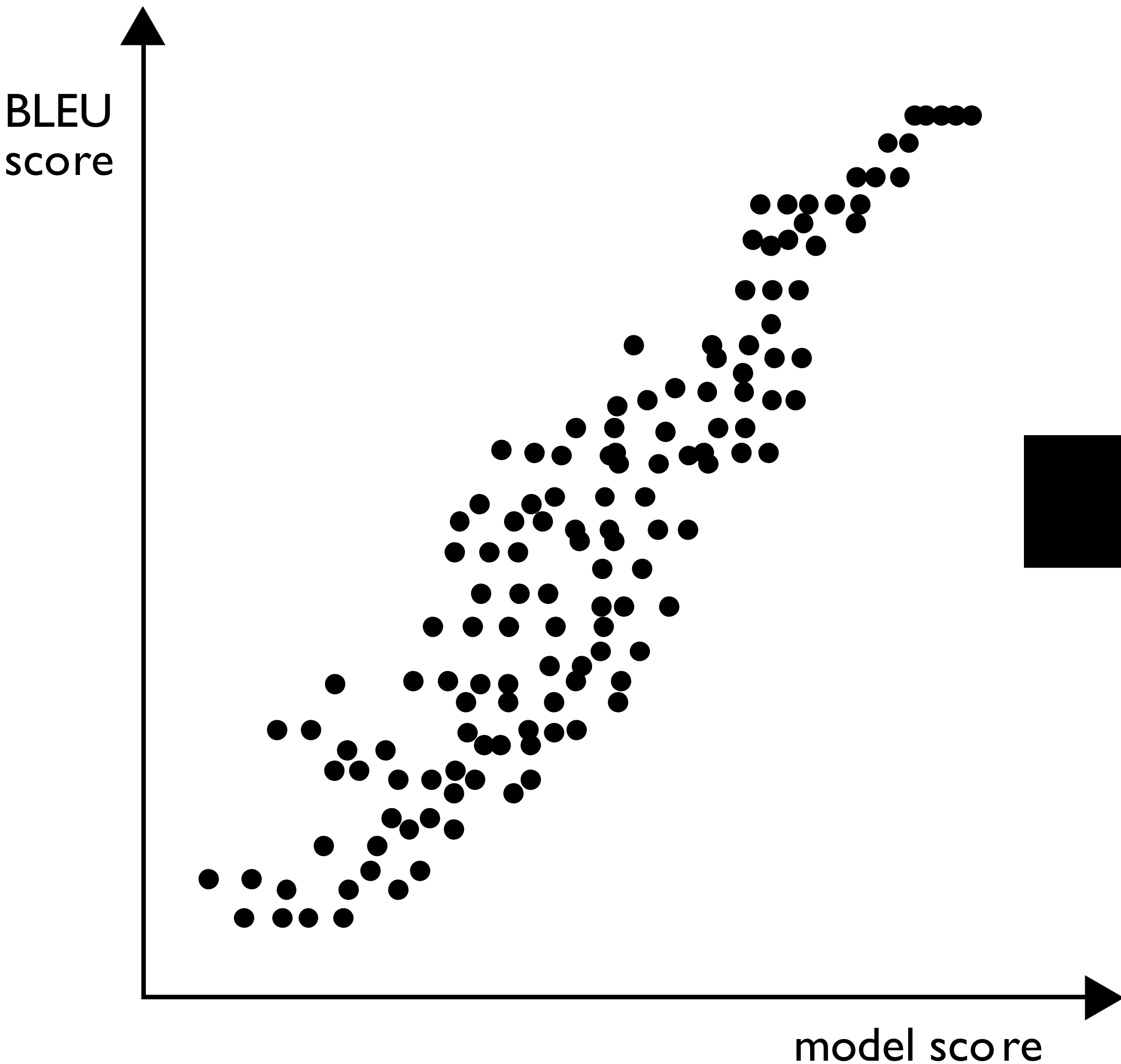
Gold standard:
African National Congress opposes sanctions against Zimbabwe



learning moves translations left or right in this plot

African National Congress 非国大　opposition 反对　sanction 制裁　Zimbabwe 津巴布韦

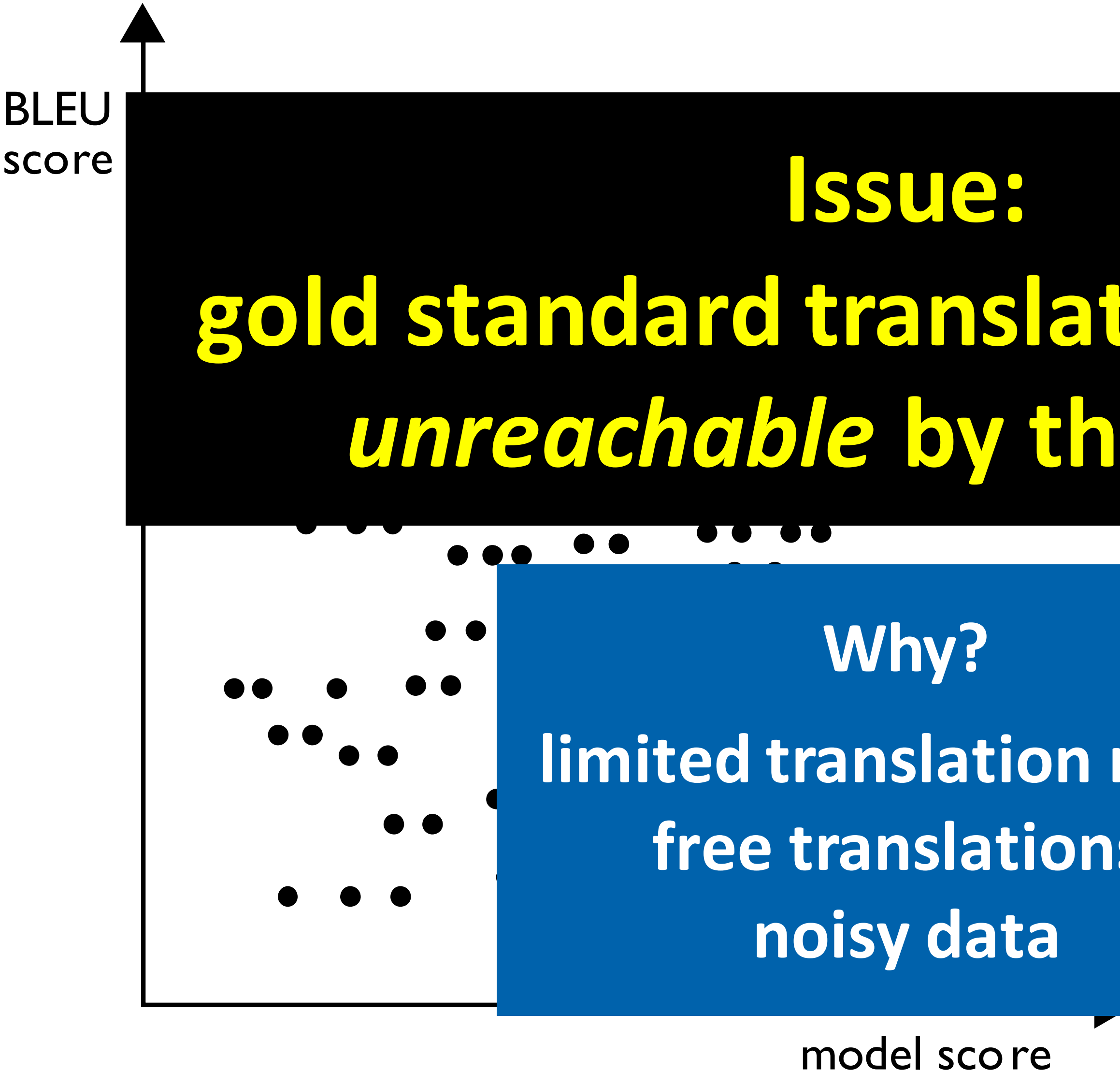
Gold standard:
African National Congress opposes sanctions against Zimbabwe



"ideal" model

African National Congress — 非国大
opposition — 反对
sanction — 制裁
Zimbabwe — 津巴布韦

Gold standard:
African National Congress opposes sanctions against Zimbabwe

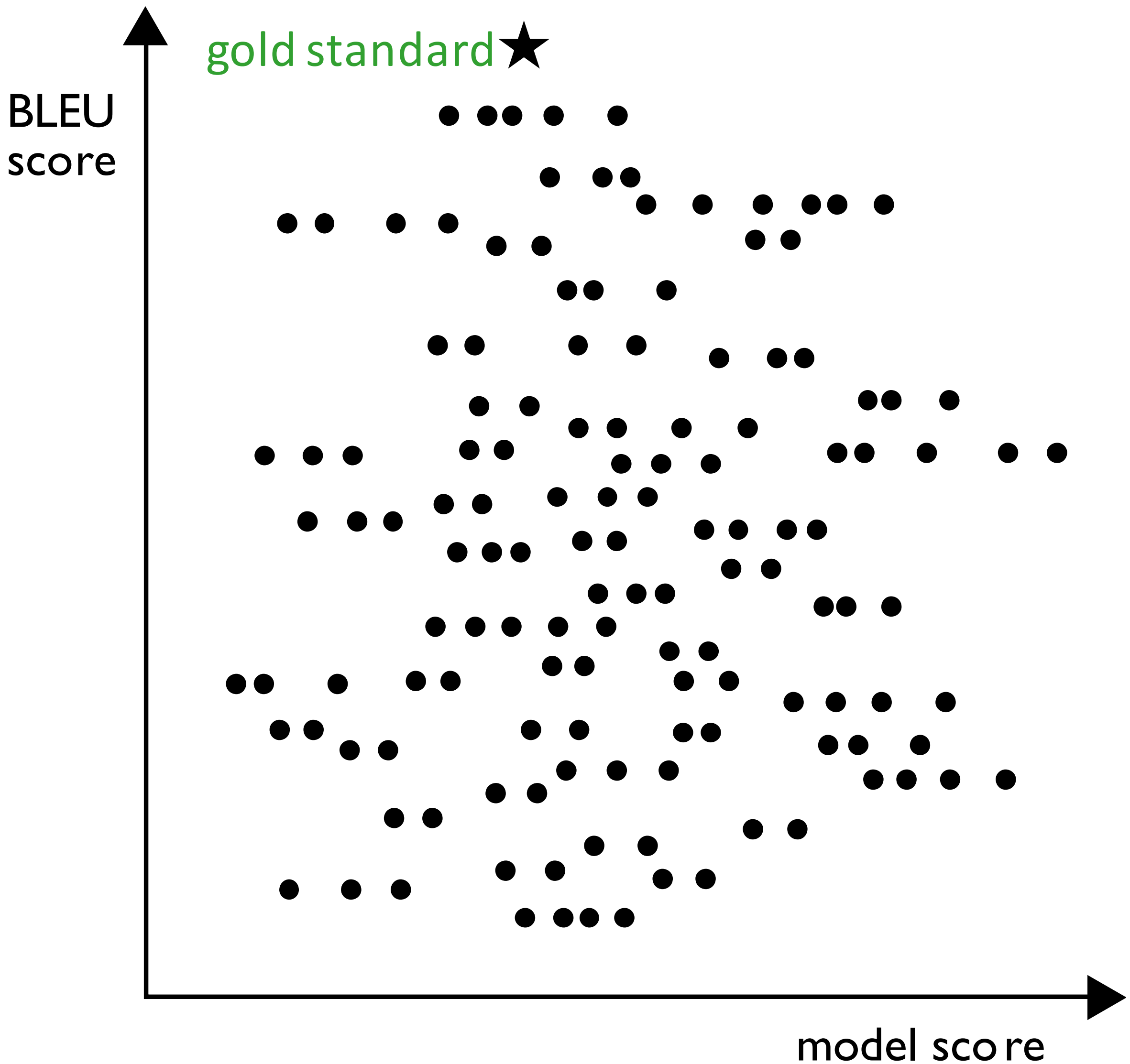


Issue:
gold standard translation is often *unreachable* by the model

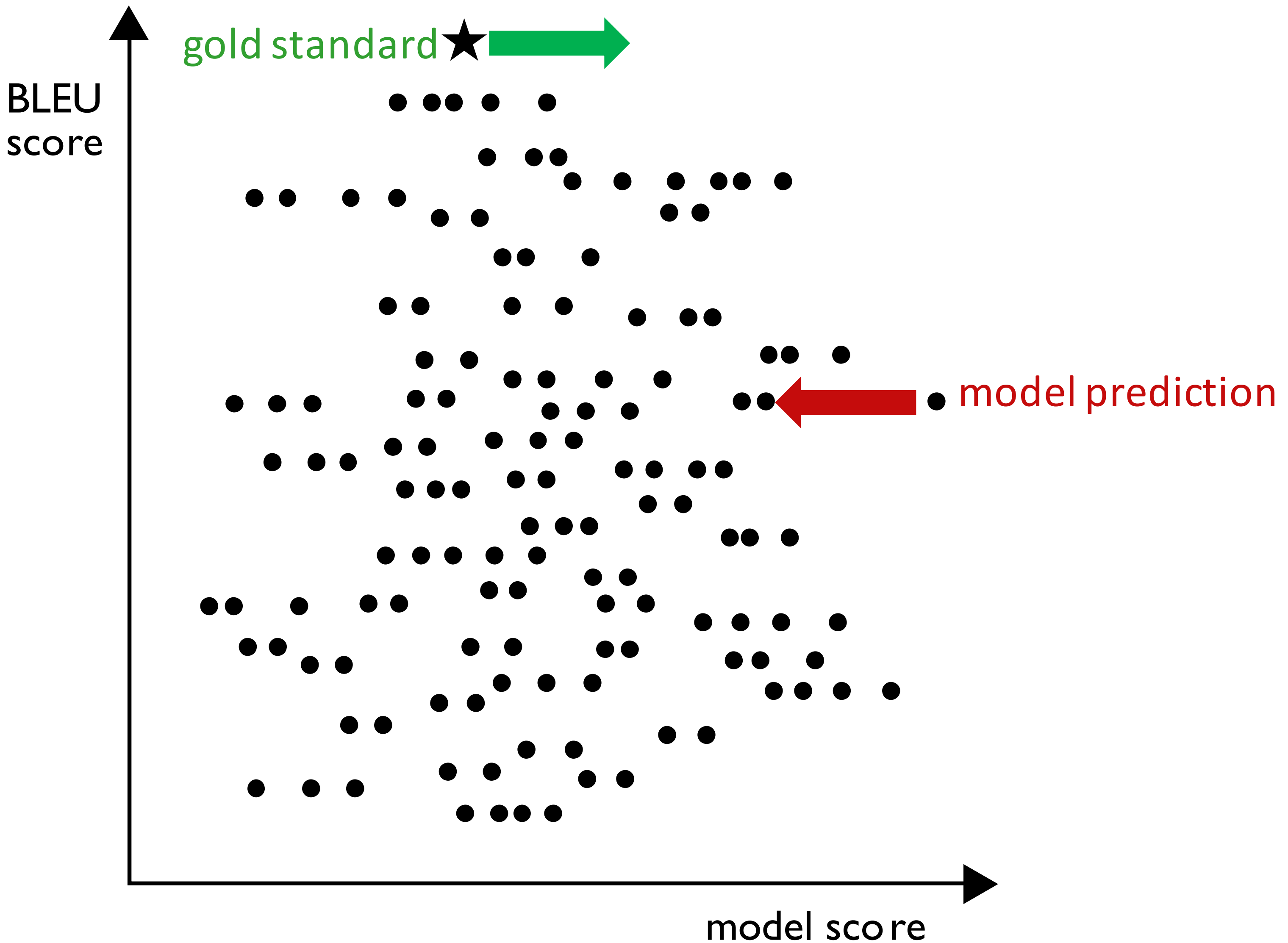
Why?

limited translation rules,
free translations,
noisy data

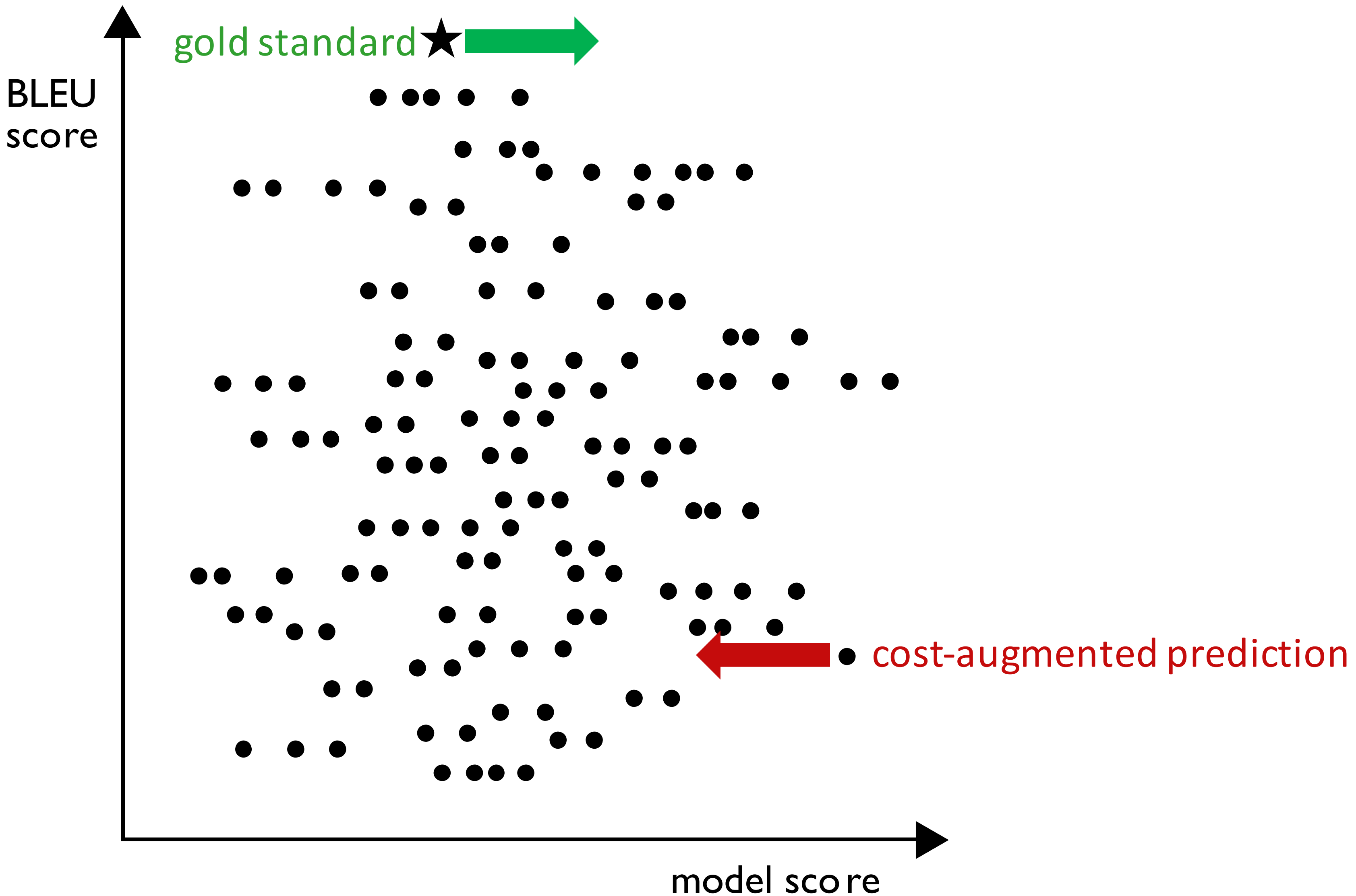
Perceptron Loss



Perceptron Loss

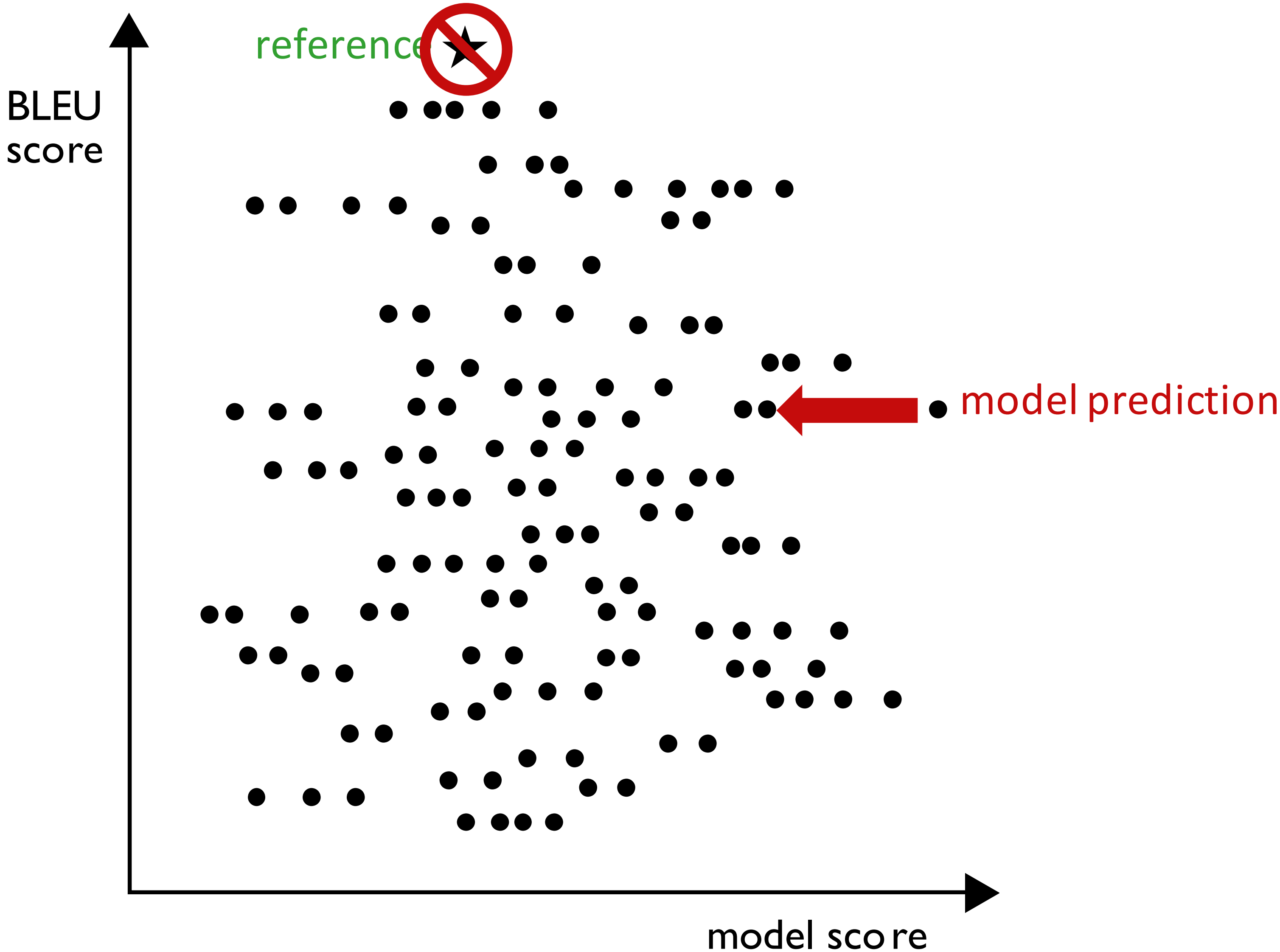


Hinge Loss

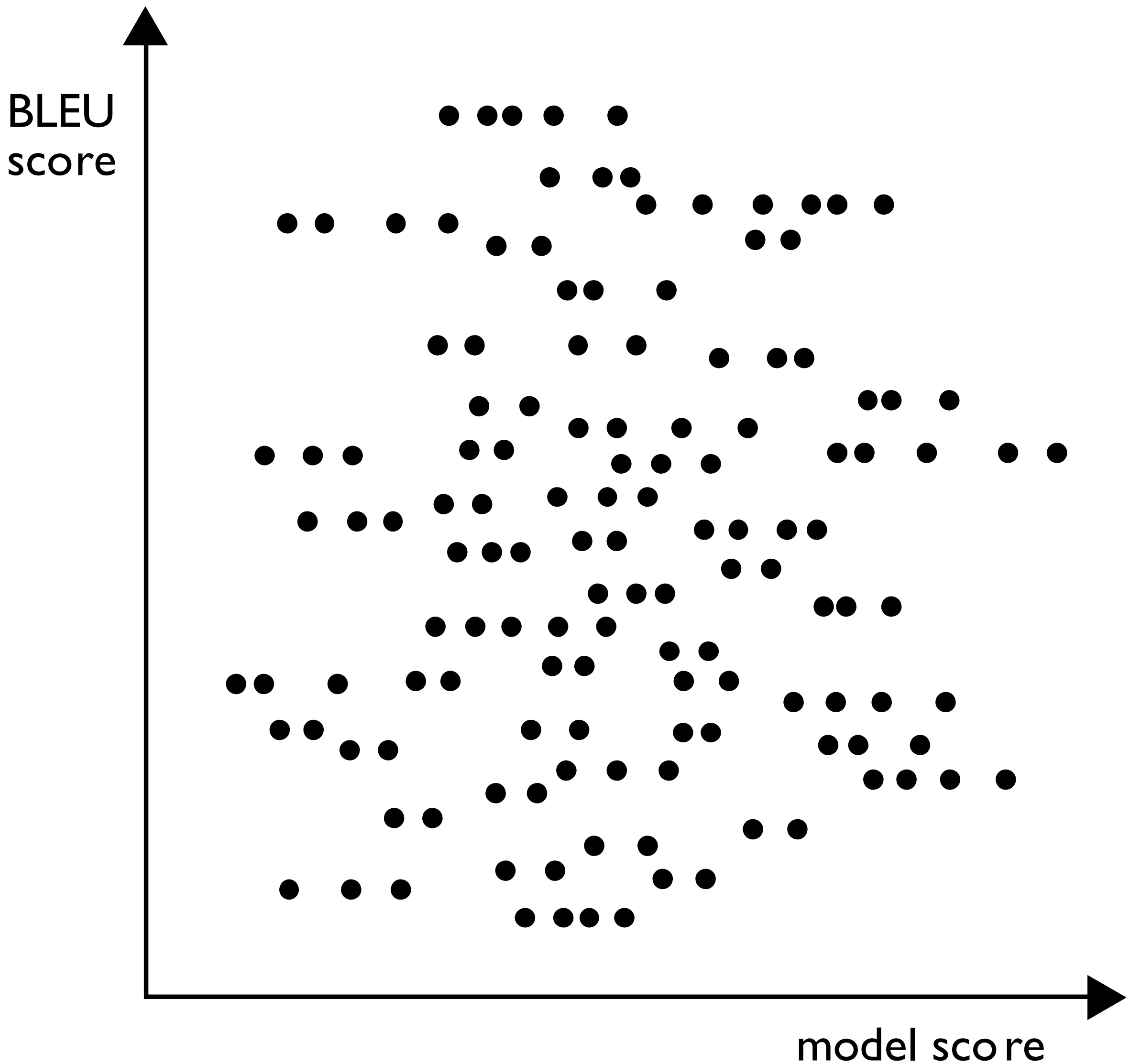


gold standard
BLEU score
cost-augmented prediction
model score

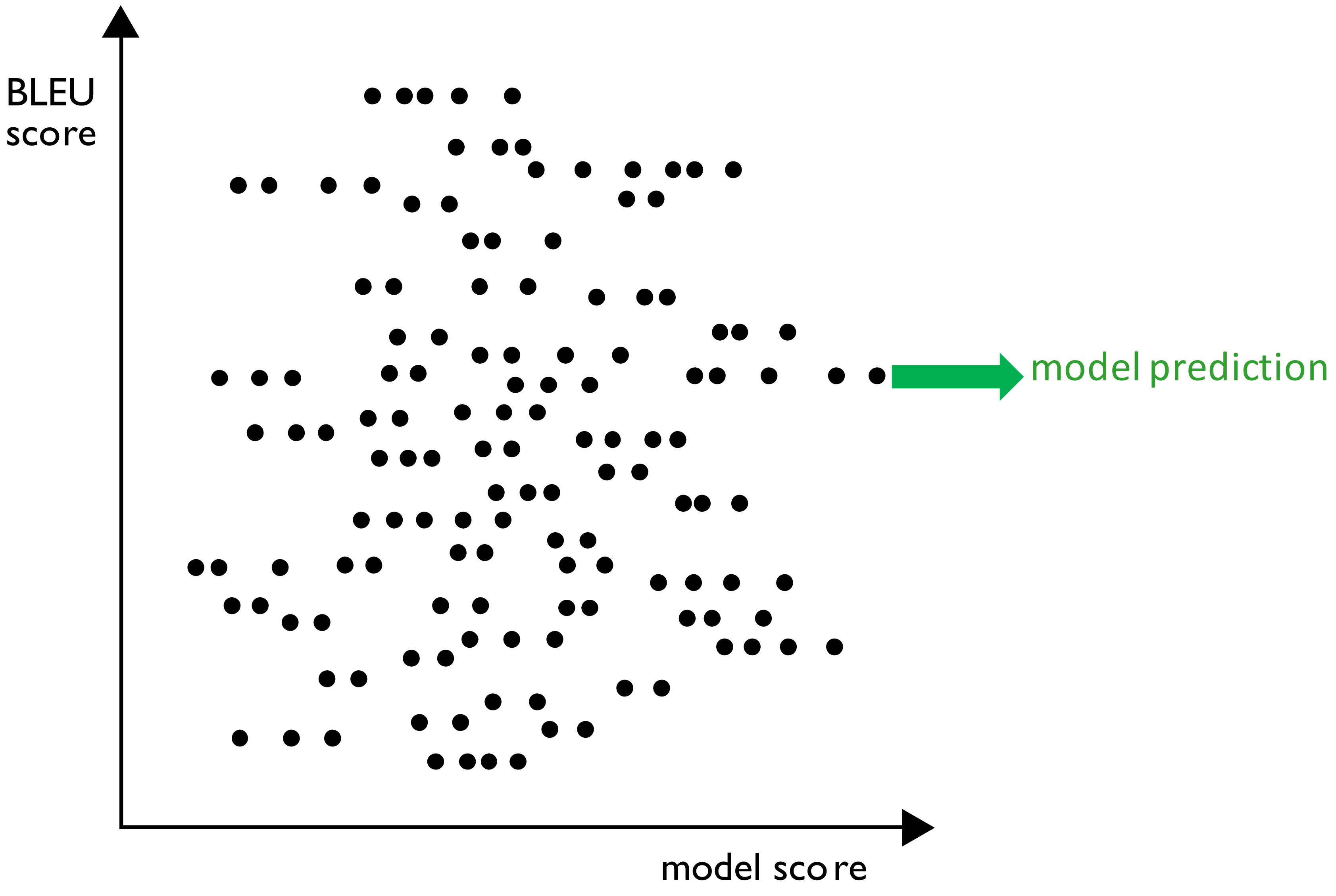
Perceptron Loss for MT?



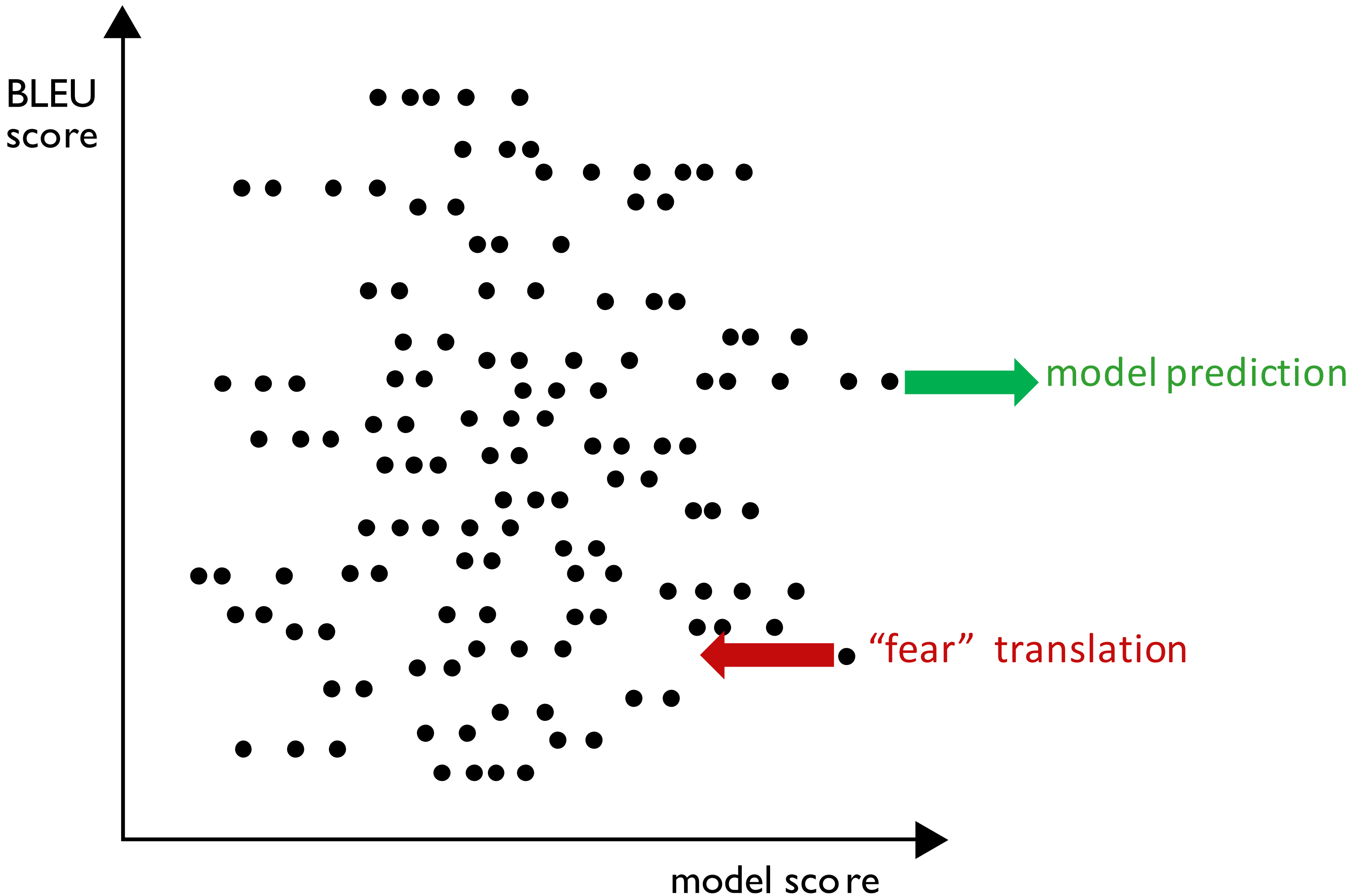
Ramp Loss Minimization



Ramp Loss Minimization

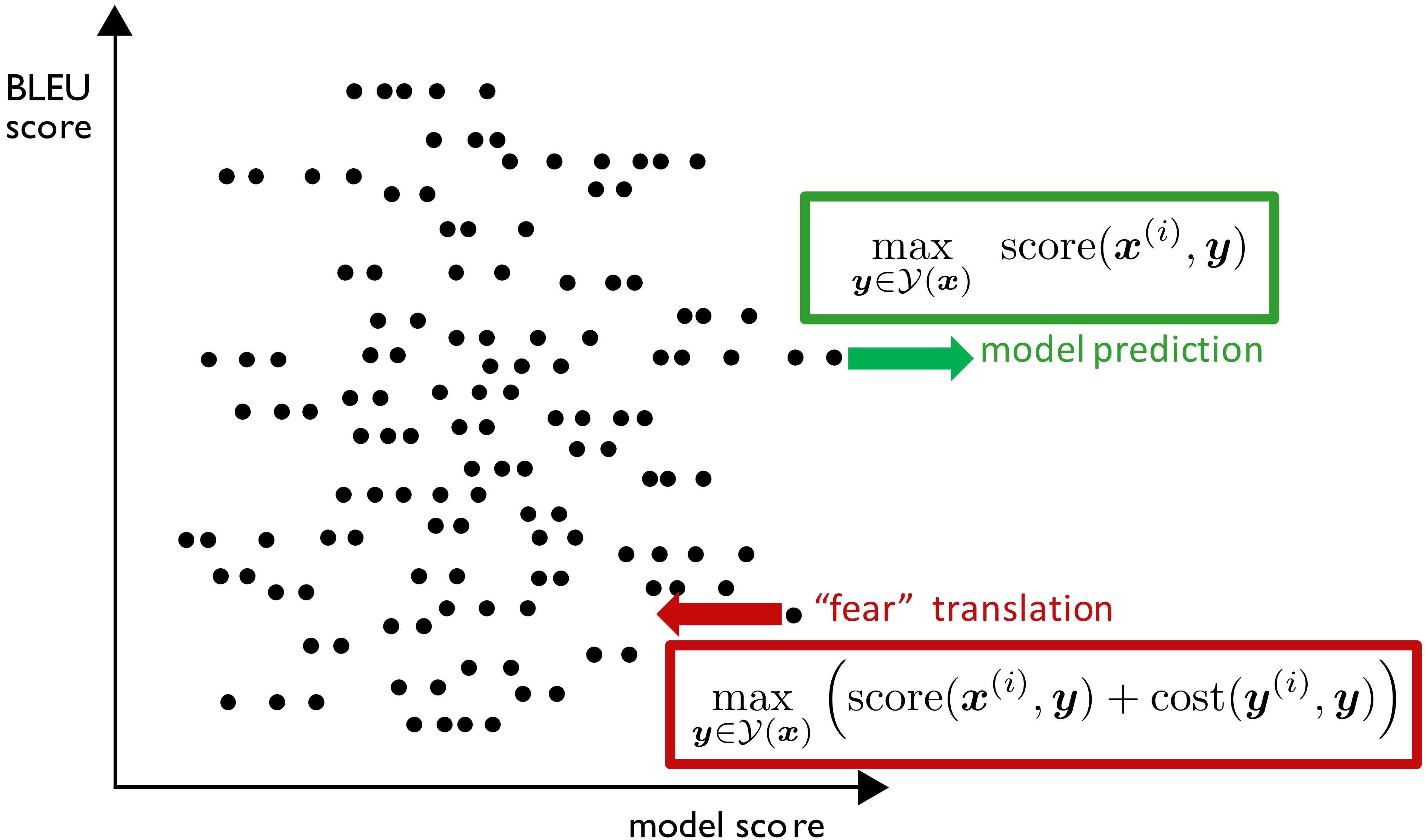


Ramp Loss Minimization



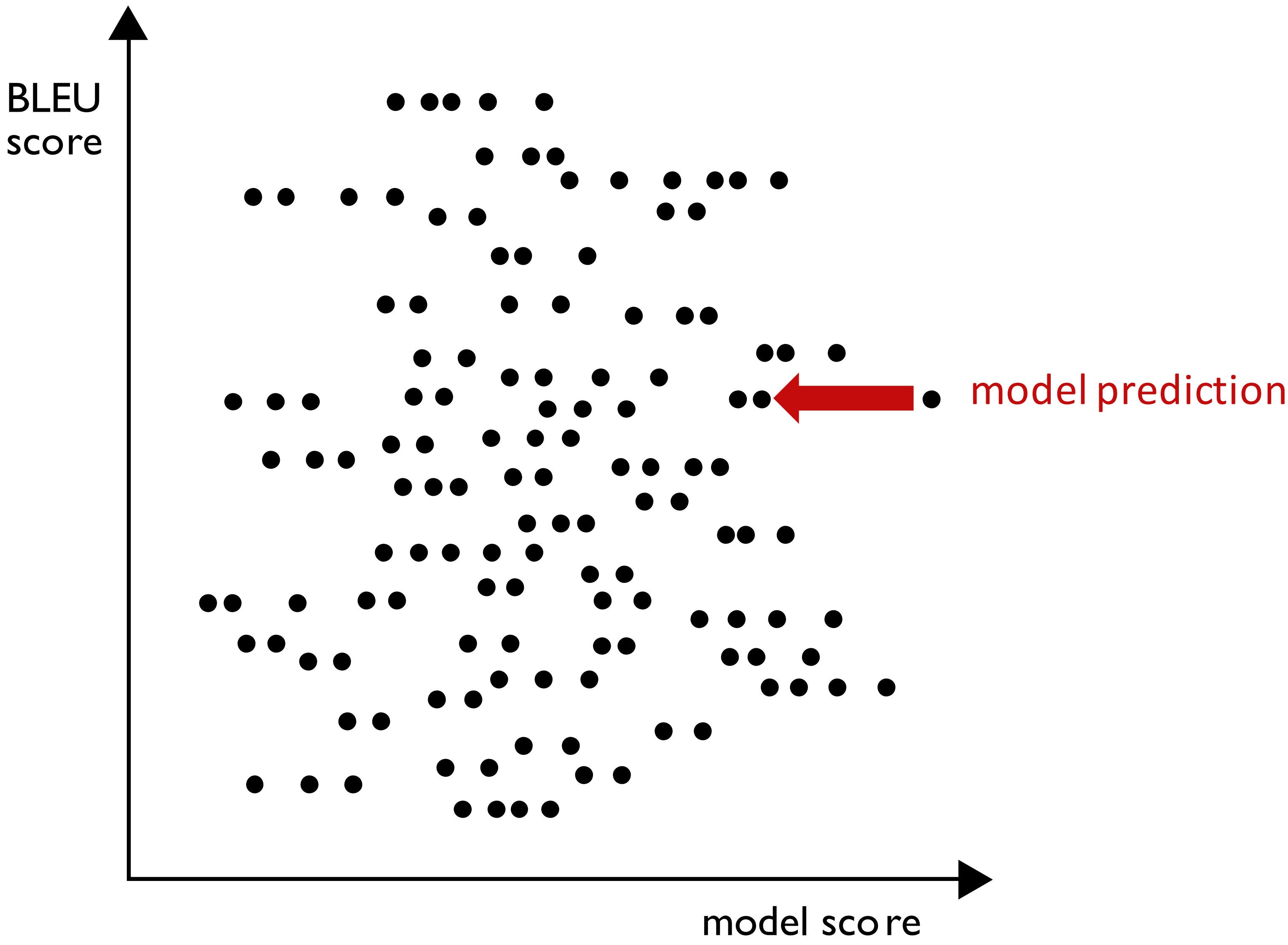
“Fear” Ramp Loss

(Do et al., 2008)



"Hope" Ramp Loss

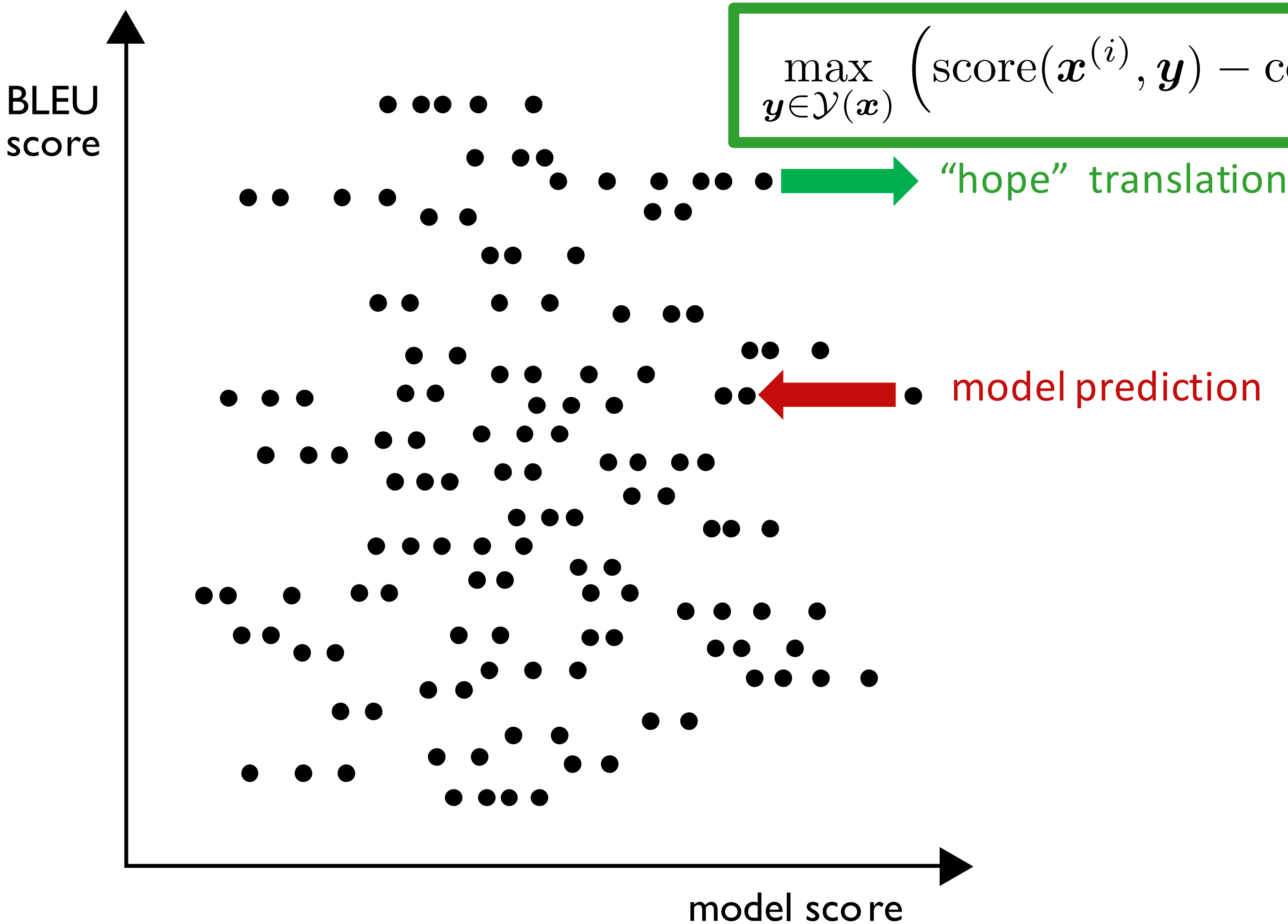
(McAllester & Keshet, 2011; Liang et al., 2006)



“Hope” Ramp Loss

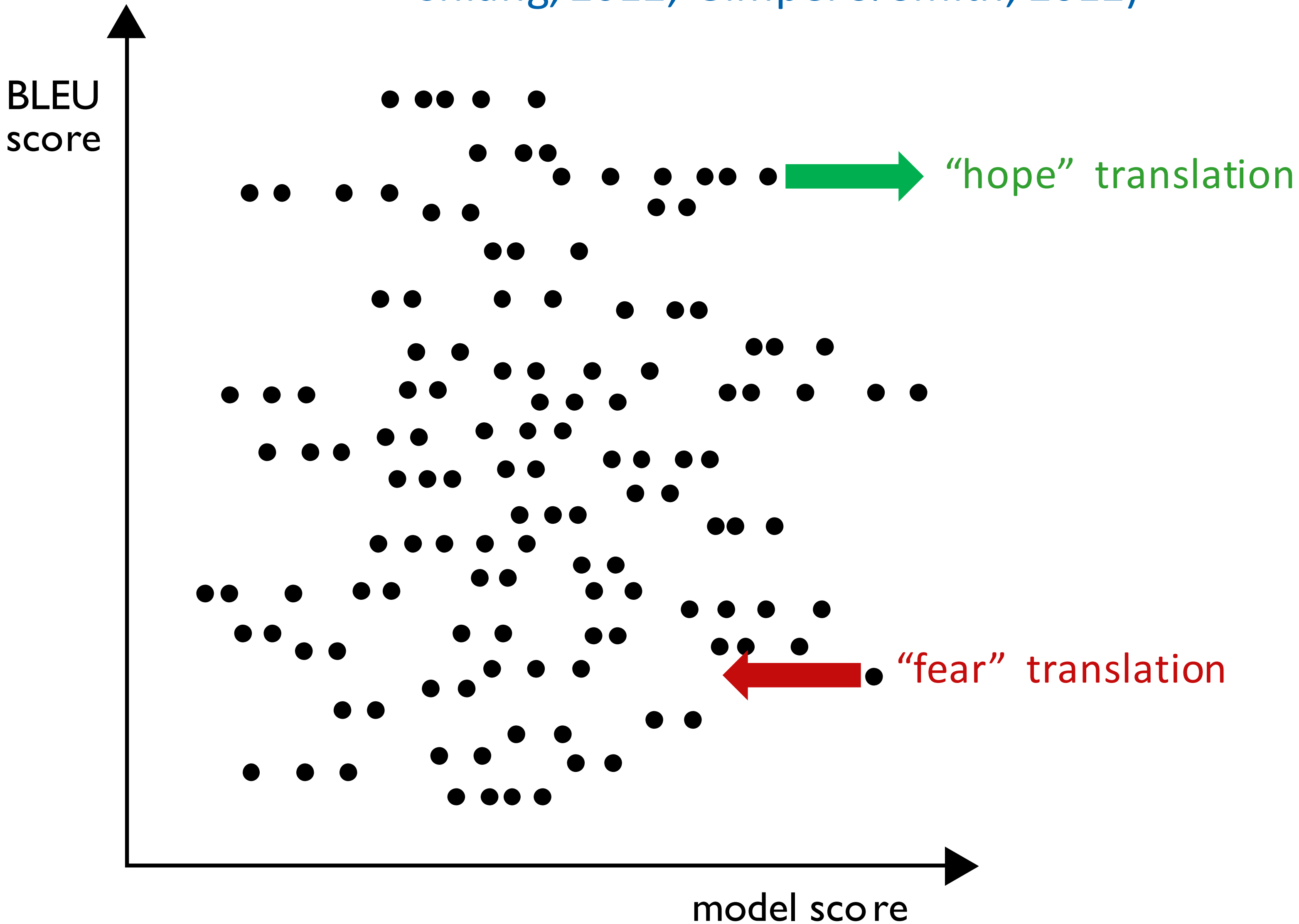
(McAllester & Keshet, 2011; Liang et al., 2006)

$$\max_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} \left(\text{score}(\boldsymbol{x}^{(i)}, \boldsymbol{y}) - \text{cost}(\boldsymbol{y}^{(i)}, \boldsymbol{y}) \right)$$



“Hope-Fear” Ramp Loss

(Chiang et al., 2008; 2009; Cherry & Foster, 2012; Chiang, 2012; Gimpel & Smith, 2012)



Experiments

(Gimpel, 2012)

averages over 8 test sets across 3 language pairs

	Moses %BLEU	Hiero %BLEU
MERT	35.9	37.0
Fear Ramp (away from bad)	34.9	34.2
Hope Ramp (toward good)	35.2	36.0
Hope-Fear Ramp (toward good + away from bad)	35.7	37.0

Why do you think that hope ramp works better than fear ramp?

I think: going away from something bad does not necessarily mean that you are going toward something good.

you might be going toward something else that's bad!

Classification Framework for Machine Translation

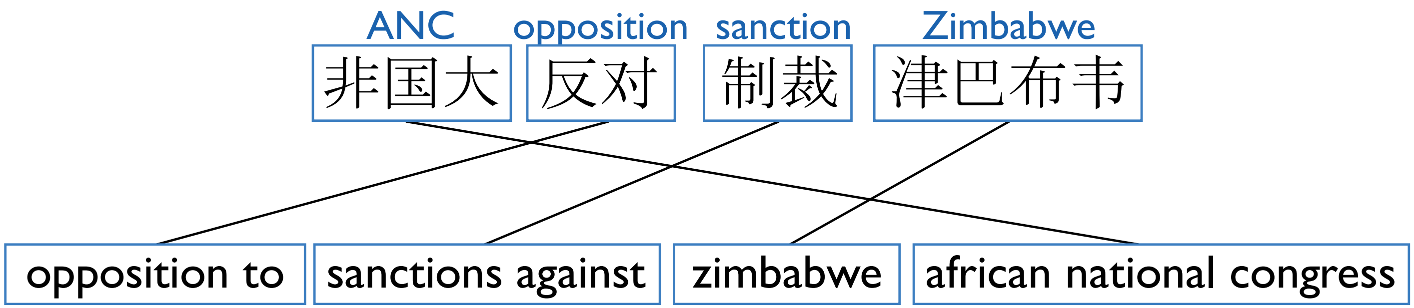
inference: solve argmax

$$\boldsymbol{y}^* = \text{classify}(\boldsymbol{x}, \boldsymbol{\theta}) = \underset{\boldsymbol{y}}{\text{argmax}} \ \text{score}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})$$

- we have a latent variable, so this becomes:

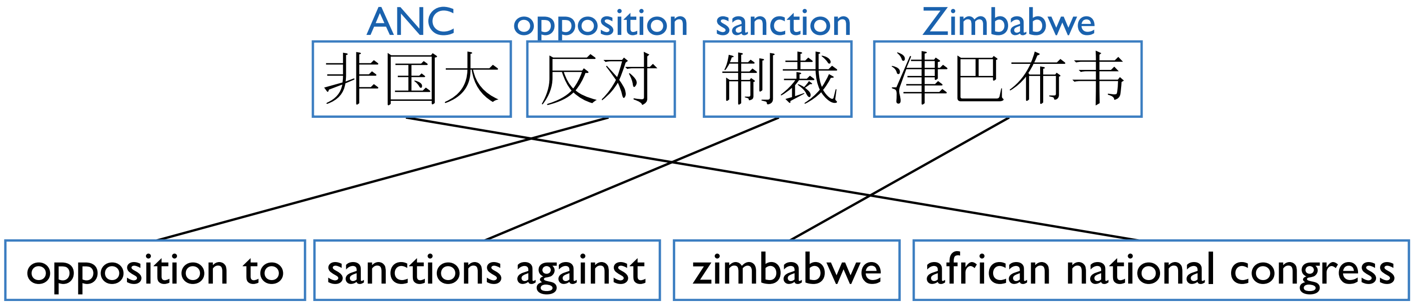
$$\langle \boldsymbol{y}^*, \boldsymbol{h}^* \rangle = \text{classify}(\boldsymbol{x}, \boldsymbol{\theta}) = \underset{\langle \boldsymbol{y}, \boldsymbol{h} \rangle}{\text{argmax}} \ \text{score}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{h}, \boldsymbol{\theta})$$

- we maximize over the latent variable AND the output!
- *h* could be word alignments, phrase segmentations/ alignments, synchronous CFG derivations, etc.



Reference: african national congress opposes sanctions against zimbabwe

- For phrase-based translation, search over:
  - Segmentations into phrases
  - Translations for each phrase
  - Orderings of the translated phrases



Reference: african national congress opposes sanctions against zimbabwe

- For phrase-based translation, search over:
  - Segmentations into phrases
  - Translations for each phrase
  - Orderings of the translated phrases

This search problem is NP-hard (Knight, 1999)

Approximate beam search is used in practice

Phrase-Based Machine Translation

Koehn et al. (2003)

African National Congress	opposition	sanction	Zimbabwe
非国大	反对	制裁	津巴布韦

Reference translation:
African National Congress opposes sanctions against Zimbabwe

Phrase-Based Machine Translation

Koehn et al. (2003)

African National Congress	opposition	sanction	Zimbabwe
非国大	反对	制裁	津巴布韦

Reference translation:
African National Congress opposes sanctions against Zimbabwe

Phrase Table

1 非国大 / African National Congress
2 反对 / opposition to
3 反对 / is opposed to
4 制裁 / sanctions
5 制裁 津巴布韦 / sanctions against Zimbabwe
…

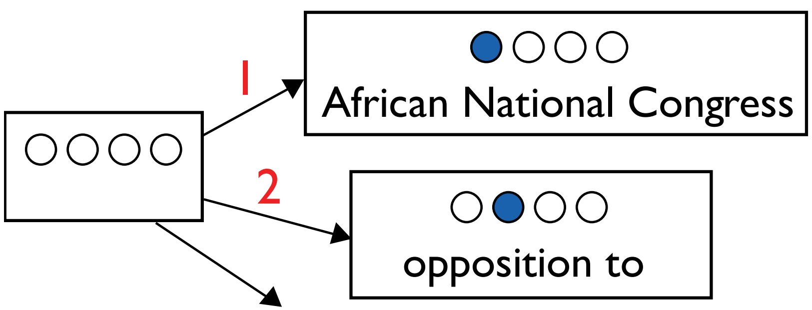
Phrase-Based Machine Translation

Koehn et al. (2003)

African National Congress	opposition	sanction	Zimbabwe
非国大	反对	制裁	津巴布韦

Reference translation:
African National Congress opposes sanctions against Zimbabwe

	Phrase Table
1	非国大 / African National Congress
2	反对 / opposition to
3	反对 / is opposed to
4	制裁 / sanctions
5	制裁 津巴布韦 / sanctions against Zimbabwe
…	



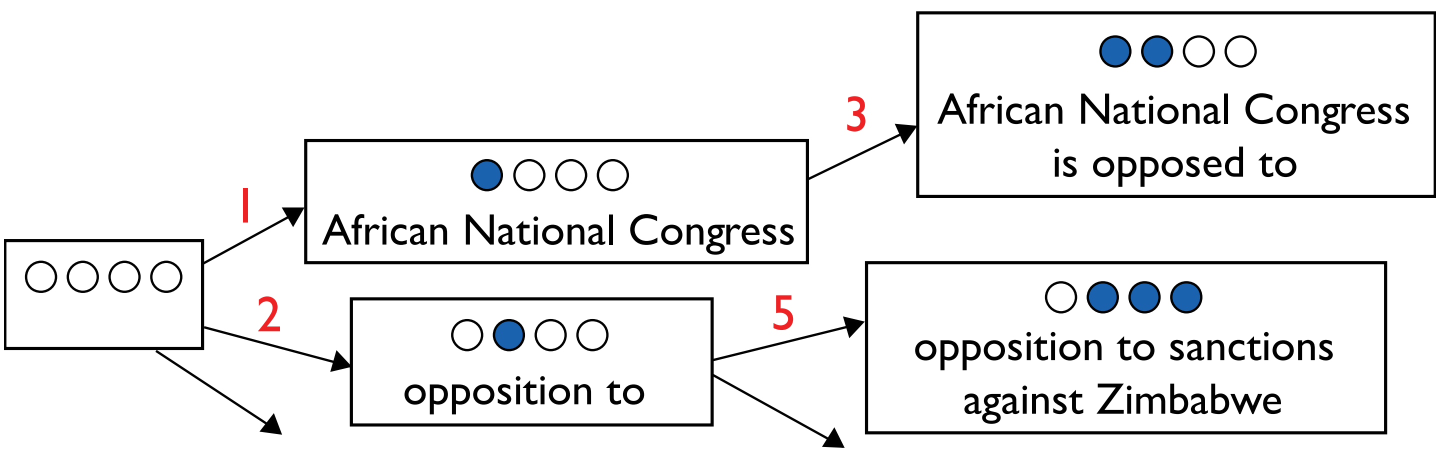
Phrase-Based Machine Translation

Koehn et al. (2003)

African National Congress	opposition	sanction	Zimbabwe
非国大	反对	制裁	津巴布韦

Reference translation:
African National Congress opposes sanctions against Zimbabwe

	Phrase Table
1	非国大 / African National Congress
2	反对 / opposition to
3	反对 / is opposed to
4	制裁 / sanctions
5	制裁 津巴布韦 / sanctions against Zimbabwe
…	



Phrase-Based Machine Translation

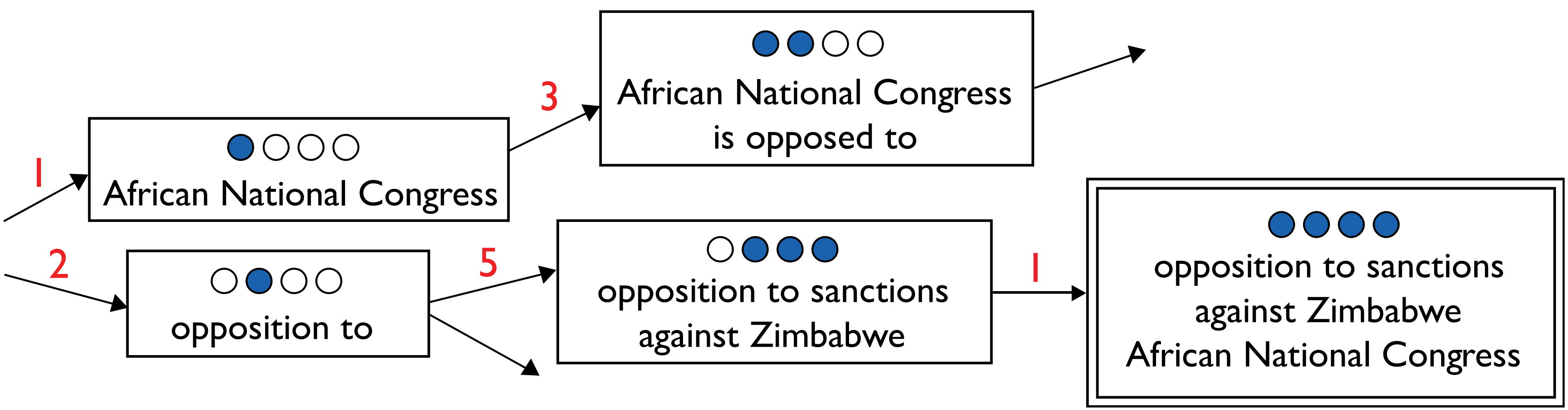
Koehn et al. (2003)

African National Congress	opposition	sanction	Zimbabwe
非国大	反对	制裁	津巴布韦

Reference translation:
African National Congress opposes sanctions against Zimbabwe

Phrase Table

1	非国大	/	African National Congress
2	反对	/	opposition to
3	反对	/	is opposed to
4	制裁	/	sanctions
5	制裁 津巴布韦	/	sanctions against Zimbabwe
…			



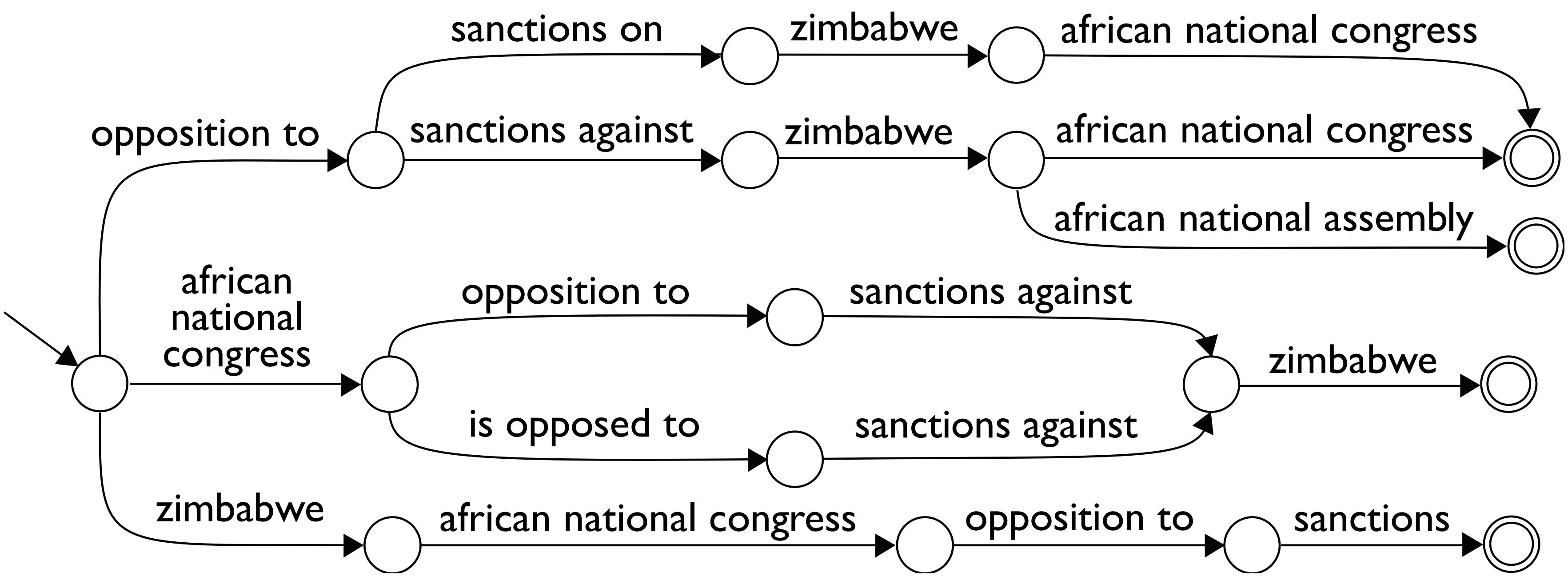
other useful inference tasks:

- find k-best translations

Rank	Score				
1	-11.8	opposition to	sanctions against	zimbabwe	african national congress
2	-12.1	african national congress	opposition to	sanctions against	zimbabwe
3	-12.4	african national congress	oppose	sanctions against	zimbabwe
4	-12.9	zimbabwe	african national congress	opposition to	sanctions
5	-13.5	opposition to	sanctions on	zimbabwe	african national congress

other useful inference tasks:

- find **phrase lattice** of translations



typical lattices contain up to 10^{80} paths!

(but not all are unique translations)

Neural Networks and Machine Translation

- current trend in MT research is to use neural networks for everything
- "neural MT" typically refers to approaches that **only** use neural networks
- but most MT systems combine traditional phrase-based models with features based on neural networks

Fast and Robust Neural Network Joint Models for Statistical Machine Translation

ACL 2014 (best paper award)

Jacob Devlin, Rabih Zbib, Zhongqiang Huang,
Thomas Lamar, Richard Schwartz, and **John Makhoul**
Raytheon BBN Technologies, 10 Moulton St, Cambridge, MA 02138, USA
{jdevlin,rzbib,zhuang,tlamar,schwartz,makhoul}@bbn.com

Abstract

Recent work has shown success in using neural network language models (NNLMs) as features in MT systems. Here, we present a novel formulation for a neural network *joint* model (NNJM), which augments the NNLM with a source context window. Our model is purely lexicalized and can be integrated into any MT decoder. We also present several variations of the NNJM which provide significant additive improvements.

Although the model is quite simple, it yields strong empirical results. On the NIST OpenMT12 Arabic-English condition, the NNJM features produce a gain of +3.0 BLEU on top of a powerful, feature-rich baseline which already includes a target-only NNLM. The NNJM features also produce a gain of +6.3 BLEU on top of a simpler baseline equivalent to Chiang's (2007) original Hiero implementation.

Fast and Robust Neural Network Joint Models for Statistical Machine Translation

ACL 2014

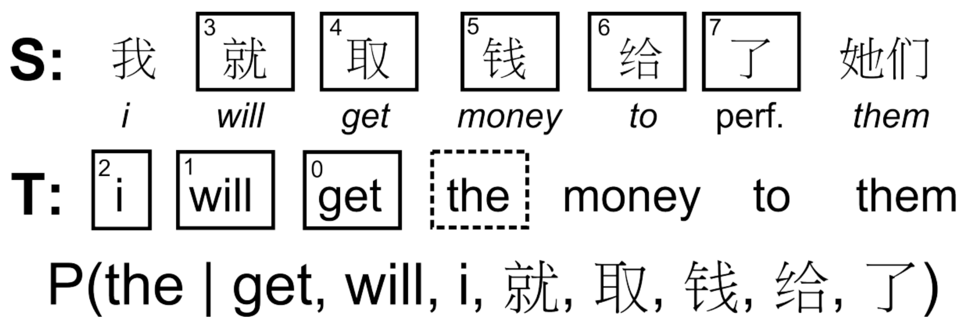


Figure 1: Context vector for target word "the", using a 3-word target history and a 5-word source window (i.e., $n = 4$ and $m = 5$). Here, "the" inherits its affiliation from "money" because this is the first aligned word to its right. The number in each box denotes the index of the word in the context vector. This indexing must be consistent across samples, but the absolute ordering does not affect results.

Fast and Robust Neural Network Joint Models for Statistical Machine Translation

ACL 2014

NIST MT12 Test		
	Ar-En	Ch-En
	BLEU	BLEU
OpenMT12 - 1st Place	49.5	32.6
OpenMT12 - 2nd Place	47.5	32.2
OpenMT12 - 3rd Place	47.4	30.8
. . .	. . .	. . .
OpenMT12 - 9th Place	44.0	27.0
OpenMT12 - 10th Place	41.2	25.7
Baseline (w/o RNNLM)	48.9	33.0
Baseline (w/ RNNLM)	49.8	33.4
+ S2T/L2R NNJM (Dec)	51.2	34.2
+ S2T NNLTM (Dec)	52.0	34.2
+ T2S NNLTM (Resc)	51.9	34.2
+ S2T/R2L NNJM (Resc)	52.2	34.3
+ T2S/L2R NNJM (Resc)	52.3	34.5
+ T2S/R2L NNJM (Resc)	52.8	34.7

Neural MT

Recurrent Continuous Translation Models

EMNLP 2013

Nal Kalchbrenner **Phil Blunsom**
Department of Computer Science
University of Oxford

Abstract

We introduce a class of probabilistic continuous translation models called Recurrent Continuous Translation Models that are purely based on continuous representations for words, phrases and sentences and do not rely on alignments or phrasal translation units. The models have a generation and a conditioning aspect. The generation of the translation is modelled with a target Recurrent Language Model, whereas the conditioning on the source sentence is modelled with a Convolutional Sentence Model. Through various experiments, we show first that our models obtain a perplexity with respect to gold translations that is $> 43\%$ lower than that of state-of-the-art alignment-based translation models.

Recurrent Continuous Translation Models

EMNLP 2013

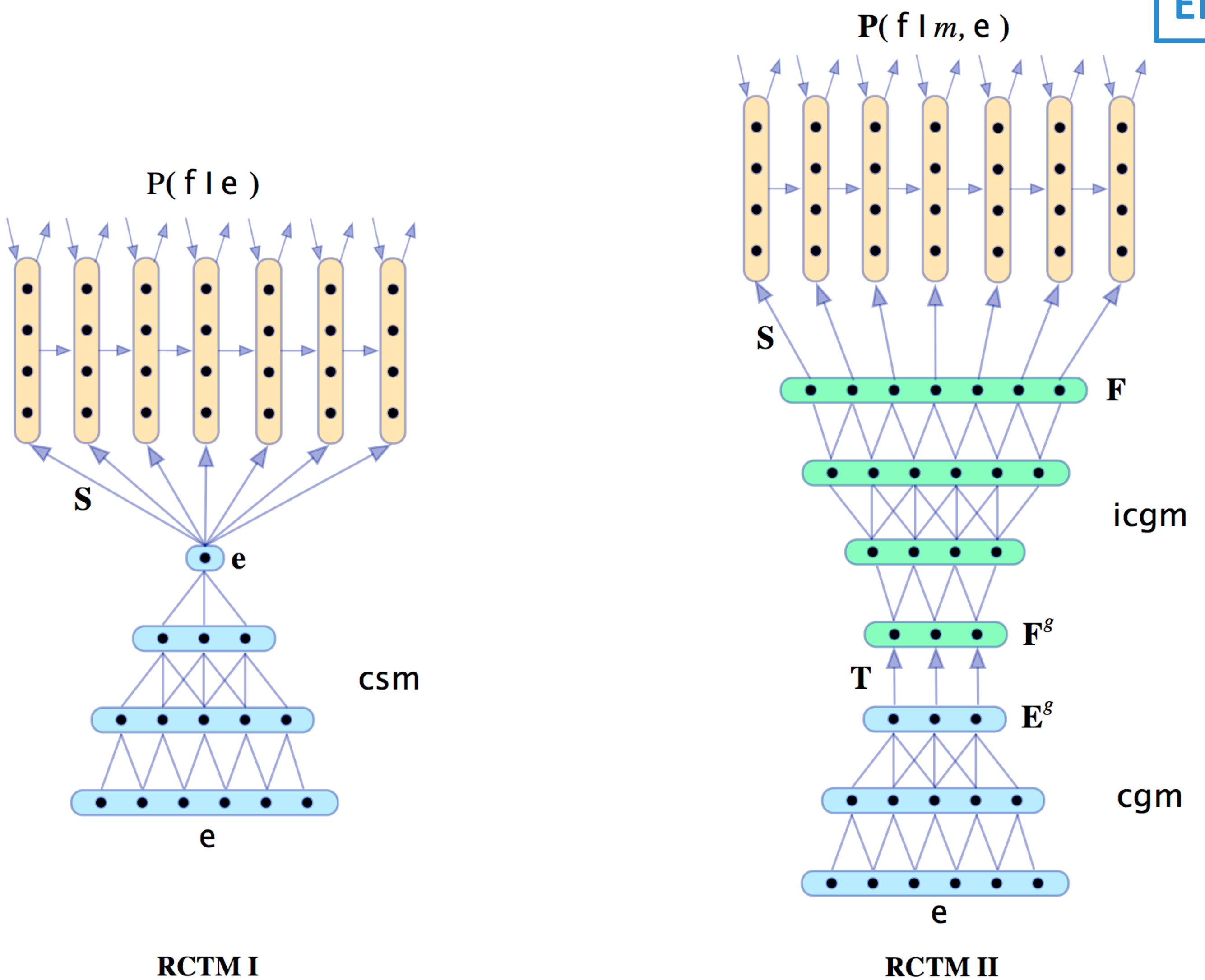


Figure 3: A graphical depiction of the two RCTMs. Arrows represent full matrix transformations while lines are vector transformations corresponding to columns of weight matrices.

Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

EMNLP 2014

Kyunghyun Cho
Bart van Merriënboer **Caglar Gulcehre**
Université de Montréal
firstname.lastname@umontreal.ca

Dzmitry Bahdanau
Jacobs University, Germany
d.bahdanau@jacobs-university.de

Fethi Bougares **Holger Schwenk**
Université du Maine, France
firstname.lastname@lium.univ-lemans.fr

Yoshua Bengio
Université de Montréal, CIFAR Senior Fellow
find.me@on.the.web

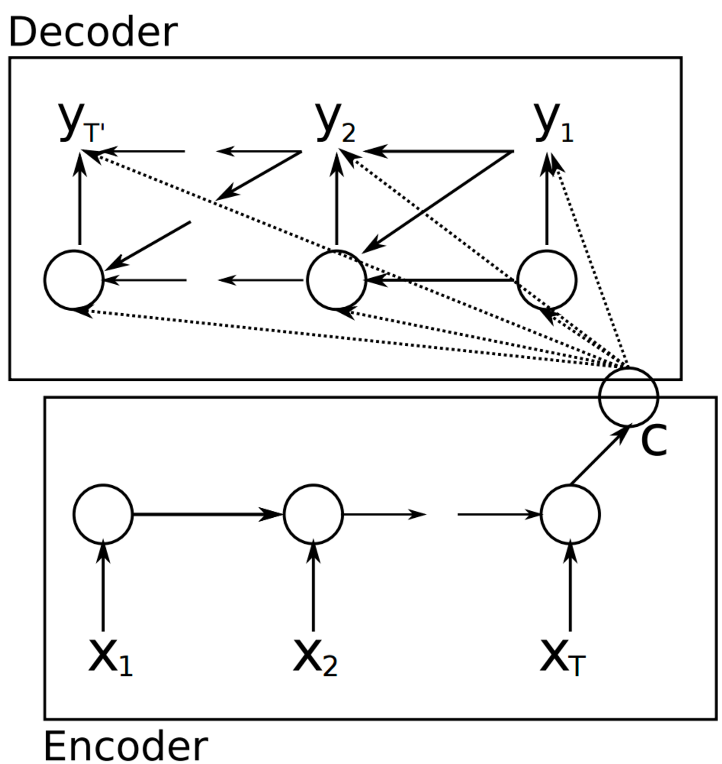
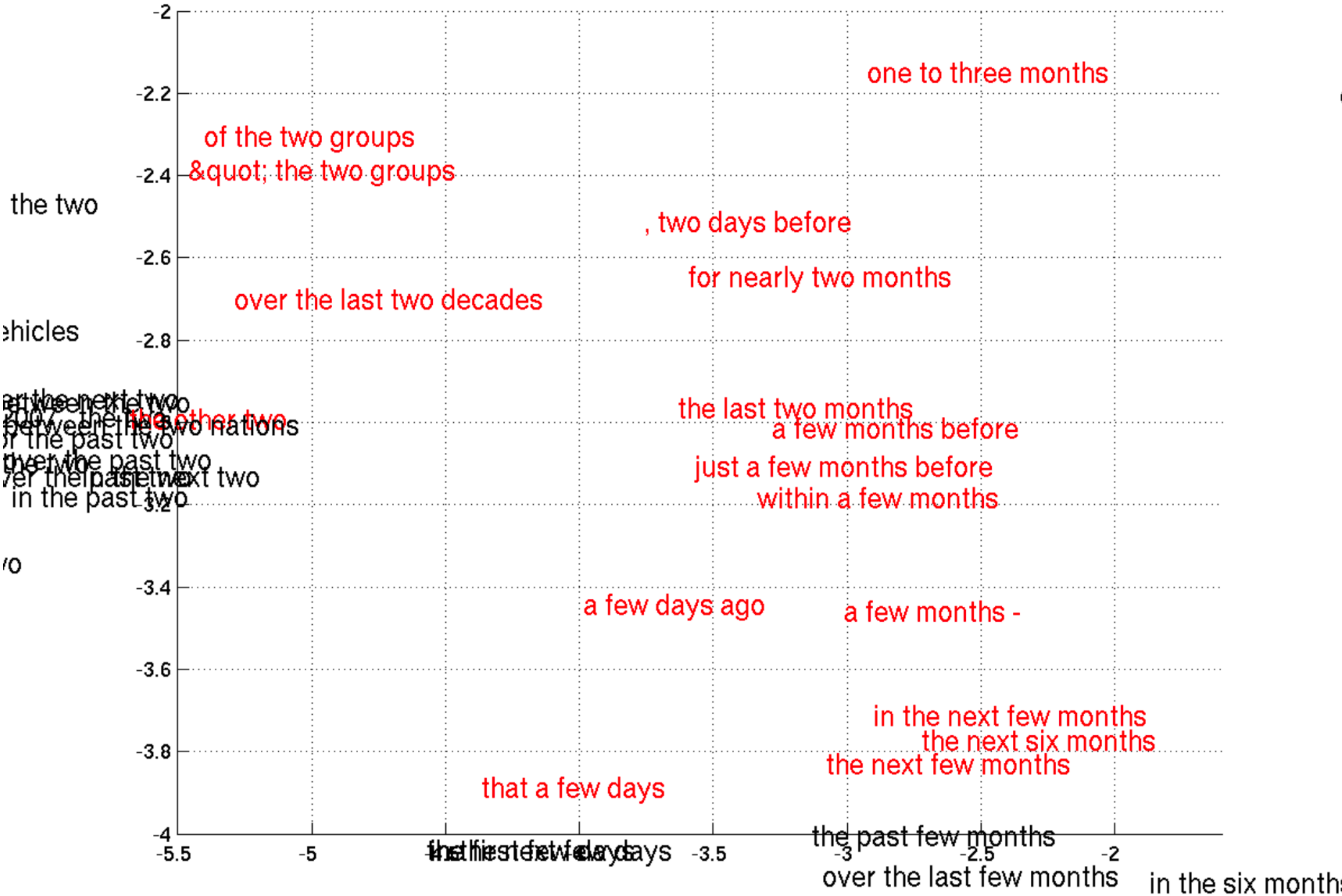


Figure 1: An illustration of the proposed RNN Encoder–Decoder.

Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

EMNLP 2014



NIPS 2014

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com

Abstract

Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences. In this paper, we present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. Our method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. Our main result is that on an English to French translation task from the WMT-14 dataset, the translations produced by the LSTM achieve a BLEU score of 34.8 on the entire test set, where the LSTM's BLEU score was penalized on out-of-vocabulary words. Additionally, the LSTM did not have difficulty on long sentences. For comparison, a phrase-based SMT system achieves a BLEU score of 33.3 on the same dataset. When we used the LSTM to rerank the 1000 hypotheses produced by the aforementioned SMT system, its BLEU score increases to 36.5, which is close to the previous state of the art. The LSTM also learned sensible phrase and sentence representations that are sensitive to word order and are relatively invariant to the active and the passive voice. Finally, we found that reversing the order of the words in all source sentences (but not target sentences) improved the LSTM's performance markedly, because doing so introduced many short term dependencies between the source and the target sentence which made the optimization problem easier.

Sequence to Sequence Learning with Neural Networks

NIPS 2014

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com

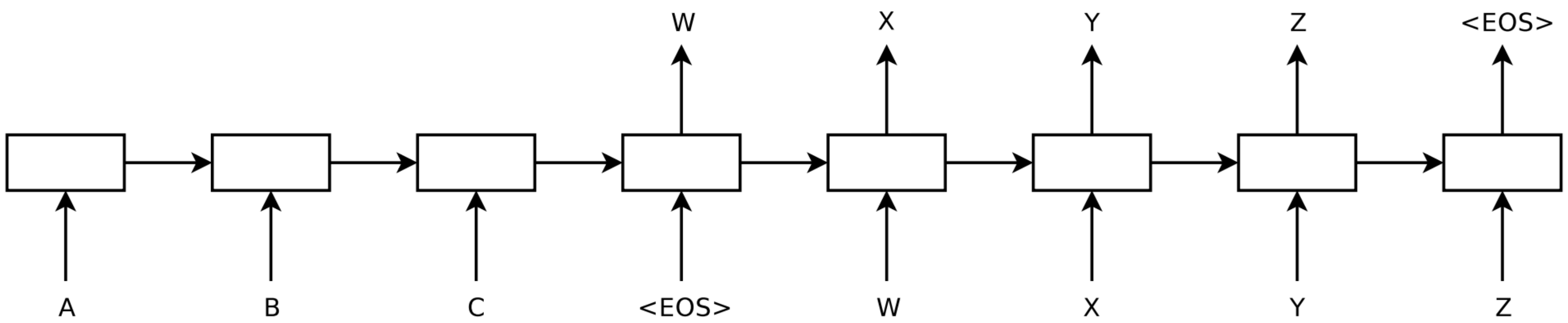


Figure 1: Our model reads an input sentence "ABC" and produces "WXYZ" as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that make the optimization problem much easier.

Sequence to Sequence Learning with Neural Networks

NIPS 2014

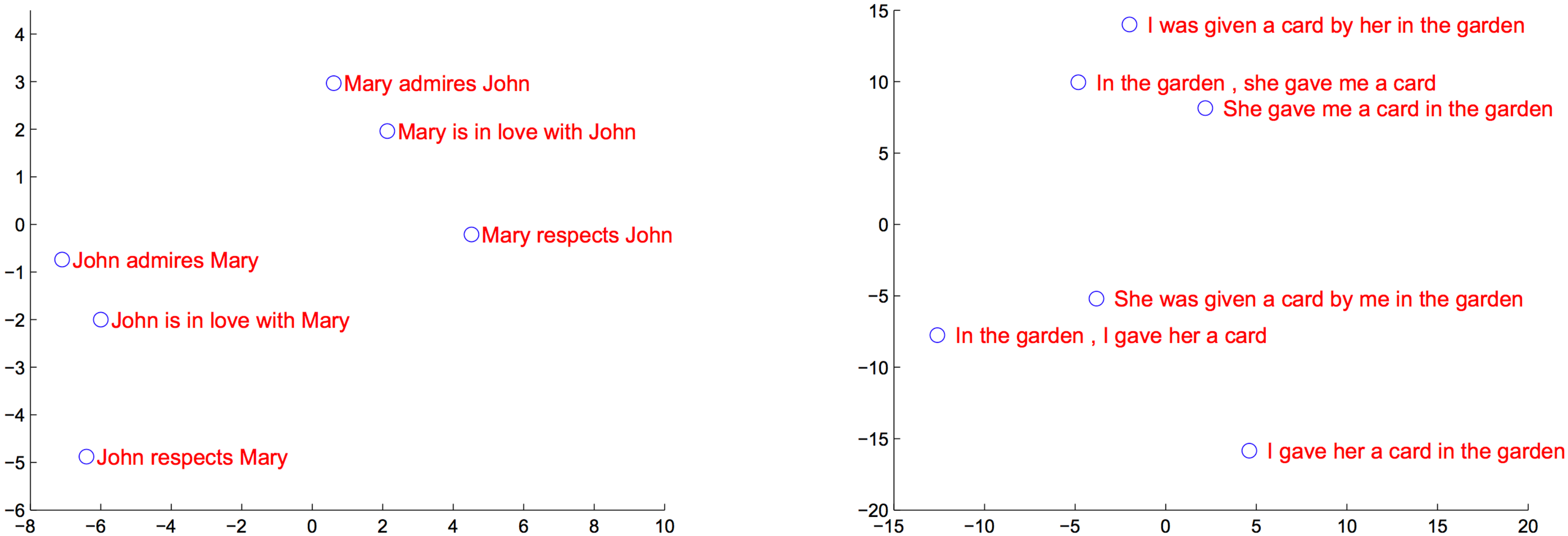


Figure 2: The figure shows a 2-dimensional PCA projection of the LSTM hidden states that are obtained after processing the phrases in the figures. The phrases are clustered by meaning, which in these examples is primarily a function of word order, which would be difficult to capture with a bag-of-words model. Notice that both clusters have similar internal structure.

Neural Machine Translation by Jointly Learning to Align and Translate

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

ICLR 2015

Abstract

Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder–decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore, qualitative analysis reveals that the (soft-)alignments found by the model agree well with our intuition.

Neural Machine Translation by Jointly Learning to Align and Translate

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

ICLR 2015

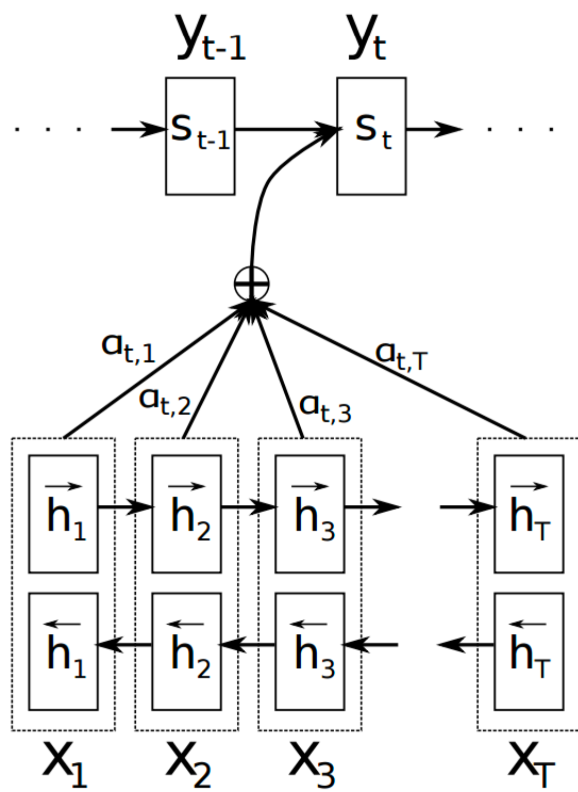
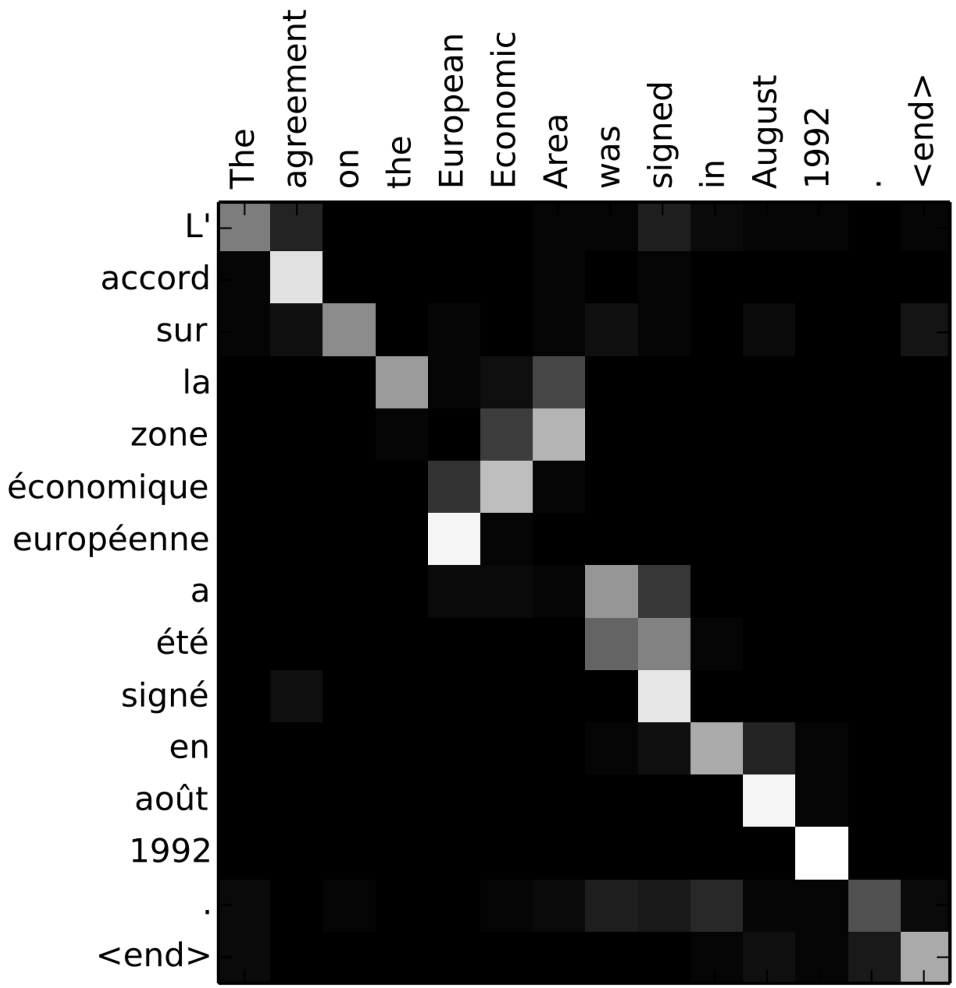


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

Neural Machine Translation by Jointly Learning to Align and Translate

ICLR 2015



Other NLP Tasks and Applications

- coreference resolution
- question answering
- summarization
- dialogue systems

Other NLP Tasks and Applications

- coreference resolution
- question answering
- summarization
- dialogue systems

Coreference Resolution

- determine which pieces of text refer to the same referent:
  - President Obama selected ten delegates after receiving recommendations from his cabinet members. They spent all day Saturday working on their recommendations for him.

Other NLP Tasks and Applications

- coreference resolution
- question answering
  - factoid question answering
  - machine comprehension
- summarization
- dialogue systems

IBM's Watson


THINK
DENKE
PENSE
$24,000
Who is Stoker?
(I FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS)
$1,000
$77,147
Who is Bram Stoker?
$ 17,973
$21,600
WHO IS BRAM STOKER?
$5600

IBM's Watson

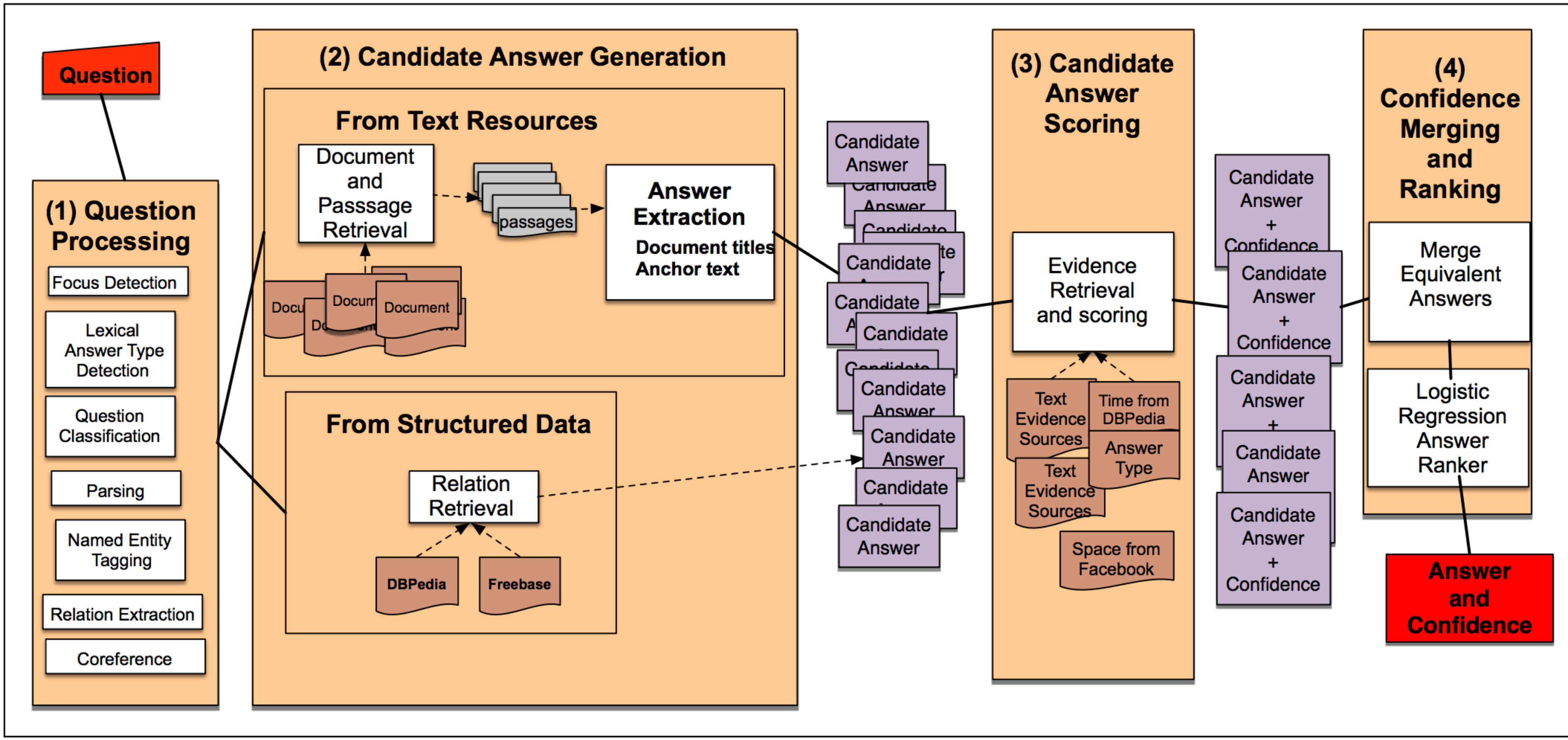


Figure 28.9 The 4 broad stages of Watson QA: (1) Question Processing, (2) Candidate Answer Generation, (3) Candidate Answer Scoring, and (4) Answer Merging and Confidence Scoring.

Classifying Questions into "Lexical Answer Types"

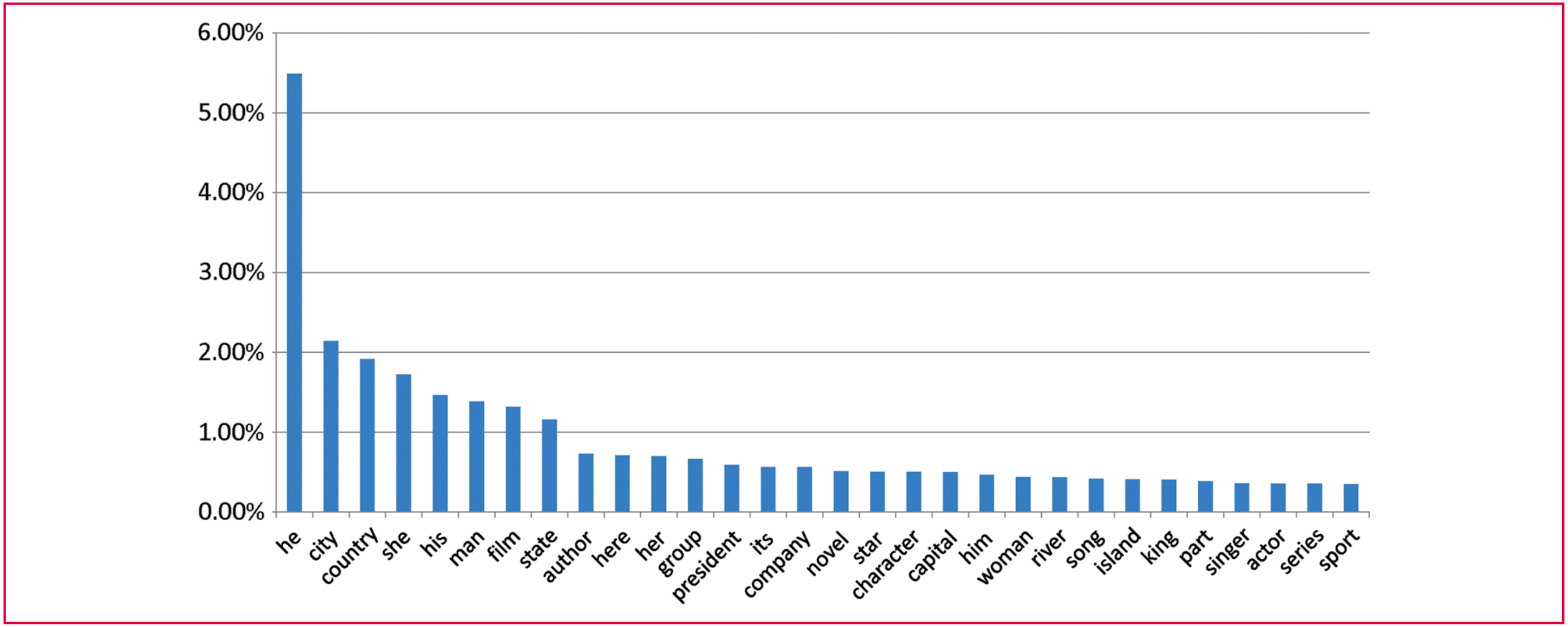


Figure 1

Distribution of the 30 most frequent lexical answer types in 20,000 Jeopardy! questions.

Other NLP Tasks and Applications

- coreference resolution
- question answering
- summarization
- dialogue systems

Automatic Summarization

- given a document, produce a summary of a provided length
- vast majority of systems are **extractive**: they extract content from the document
  - this is safer, since the document is presumably grammatical
  - but this limits applicability
- some work, especially recently, that tries to do **abstractive** summarization
  - typically based on intermediate semantic representations or neural networks

Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference

Ani Nenkova
Columbia University
1214 Amsterdam Ave
New York, NY 10027
ani@cs.columbia.edu

AAAI 2005

baseline = take first 100 words of document

regarding the first two years of DUC:

Both years, none of the systems outperforms the baseline (and the systems as a group do not outperform the baseline) and in fact the baseline has better coverage than most of the automatic systems (see the first row in table 1). It has often been noted that this baseline is indeed quite strong, due to journalistic convention for putting the most important part of an article in the initial paragraphs. But the fact that human summarizers (with the exception of F and J) significantly outperform the baseline shows that the task is meaningful and that better-than-baseline performance is possible. The

Machine Comprehension

Can a machine read a document and answer questions about it?

MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text

Matthew Richardson
Microsoft Research
One Microsoft Way
Redmond, WA 98052
mattri@microsoft.com

Christopher J.C. Burges
Microsoft Research
One Microsoft Way
Redmond, WA 98052
cburges@microsoft.com

Erin Renshaw
Microsoft Research
One Microsoft Way
Redmond, WA 98052
erinren@microsoft.com

Abstract

We present MCTest, a freely available set of stories and associated questions intended for research on the machine comprehension of text. Previous work on machine comprehension (e.g., semantic modeling) has made great strides, but primarily focuses either on limited-domain datasets, or on solving a more re-

disciplines are focused on this problem: for example, information extraction, relation extraction, semantic role labeling, and recognizing textual entailment. Yet these techniques are necessarily evaluated individually, rather than by how much they advance us towards the end goal. On the other hand, the goal of semantic parsing is the machine comprehension of text (MCT), yet its evaluation requires adherence to a specific knowledge repre-

MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text

Matthew Richardson
Microsoft Research
One Microsoft Way
Redmond, WA 98052
mattri@microsoft.com

Christopher J.C. Burges
Microsoft Research
One Microsoft Way
Redmond, WA 98052
cburges@microsoft.com

Erin Renshaw
Microsoft Research
One Microsoft Way
Redmond, WA 98052
erinren@microsoft.com

- 660 fictional stories, written at a 4th grade reading level
- 4 multiple choice questions per story

research on the machine comprehension of text. Previous work on machine comprehension (e.g., semantic modeling) has made great strides, but primarily focuses either on limited-domain datasets, or on solving a more re-

evaluated individually, rather than by how much they advance us towards the end goal. On the other hand, the goal of semantic parsing is the machine comprehension of text (MCT), yet its evaluation requires adherence to a specific knowledge repre-

Once there was a boy named Fritz who loved to draw. He drew everything. In the morning, he drew a picture of his cereal with milk. His papa said, “Don’t draw your cereal. Eat it!”

After school, Fritz drew a picture of his bicycle. His uncle said, “Don't draw your bicycle. Ride it!”

…

Once there was a boy named Fritz who loved to draw. He drew everything. In the morning, he drew a picture of his cereal with milk. His papa said, “Don’t draw your cereal. Eat it!”

After school, Fritz drew a picture of his bicycle. His uncle said, “Don't draw your bicycle. Ride it!”

…

What did Fritz draw first?

A) the toothpaste

B) his mama

C) cereal and milk

D) his bicycle

Once there was a boy named Fritz who loved to draw. He drew everything. In the morning, he drew a picture of his cereal with milk. His papa said, “Don’t draw your cereal. Eat it!”

After school, Fritz drew a picture of his bicycle. His uncle said, “Don't draw your bicycle. Ride it!”

…

What did Fritz draw first?

A) the toothpaste

B) his mama

C) cereal and milk

D) his bicycle

Once there was a boy named Fritz who loved to draw. He drew **everything**. In the morning, he drew a picture of his cereal with milk. His papa said, “Don’t draw your cereal. Eat it!”

After school, Fritz drew a picture of his bicycle. His uncle said, “Don't draw your bicycle. Ride it!”

…

What did Fritz draw first?

A) the toothpaste

B) his mama

C) cereal and milk

D) his bicycle

E) everything

- Some questions are much easier
- Simple word overlap baseline gets 63% correct

James the Turtle was always getting in trouble.

…

What is the name of the trouble making turtle?

A) Fries
B) Pudding
C) James
D) Jane

MCTest Leaderboard

institution	year	accuracy (%)
TTI-Chicago	2015	69.9
Carnegie Mellon	2015	67.8
University College London	2015	66.0
MIT	2015	63.8
Microsoft Research	2013	63.3

Our system uses several types of automatic linguistic analysis:

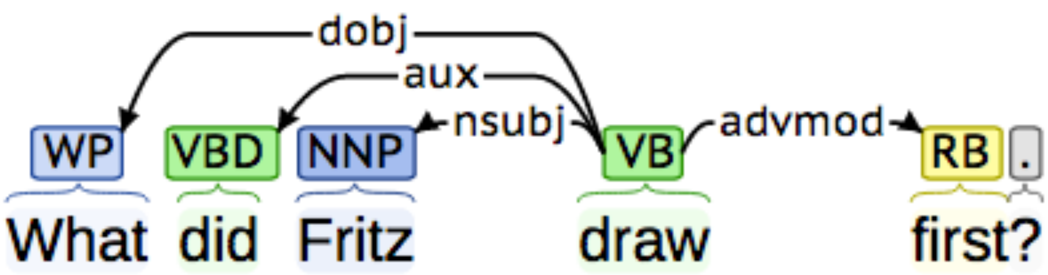
- dependency parsing
- frame semantic parsing
- coreference
- word embeddings

Our system uses several types of automatic linguistic analysis:

- **dependency parsing**

Our system uses several types of automatic linguistic analysis:

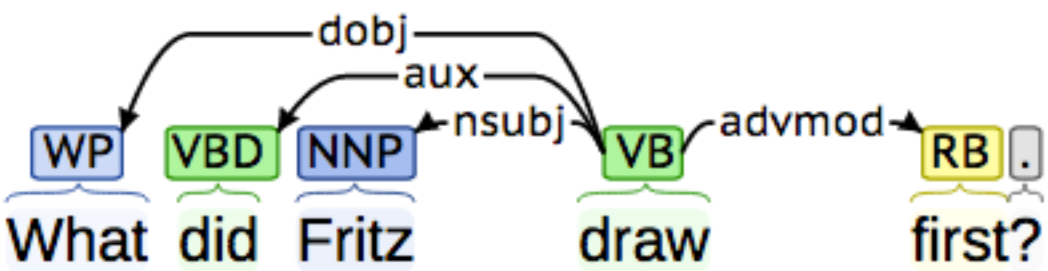
– **dependency parsing**



output of Stanford dependency parser

Our system uses several types of automatic linguistic analysis:

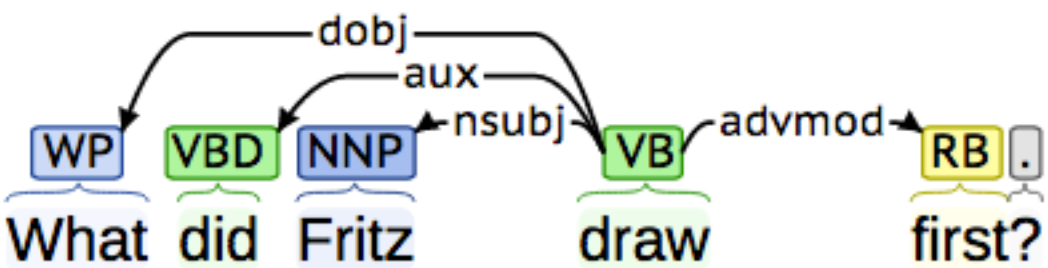
– **dependency parsing**



Fritz draw X first

Our system uses several types of automatic linguistic analysis:

– **dependency parsing**



Fritz draw X first

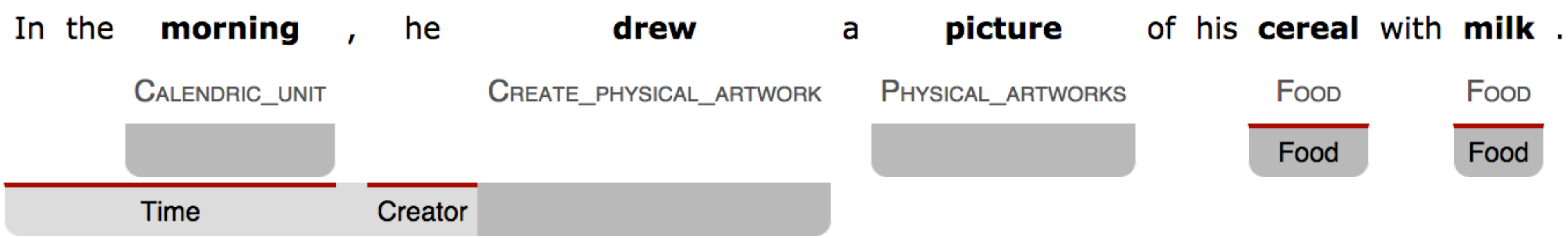
Fritz draw the toothpaste first
Fritz draw his mama first
Fritz draw cereal and milk first
Fritz draw his bicycle first

Our system uses several types of automatic linguistic analysis:

- dependency parsing
- **frame semantic parsing**

Our system uses several types of automatic linguistic analysis:

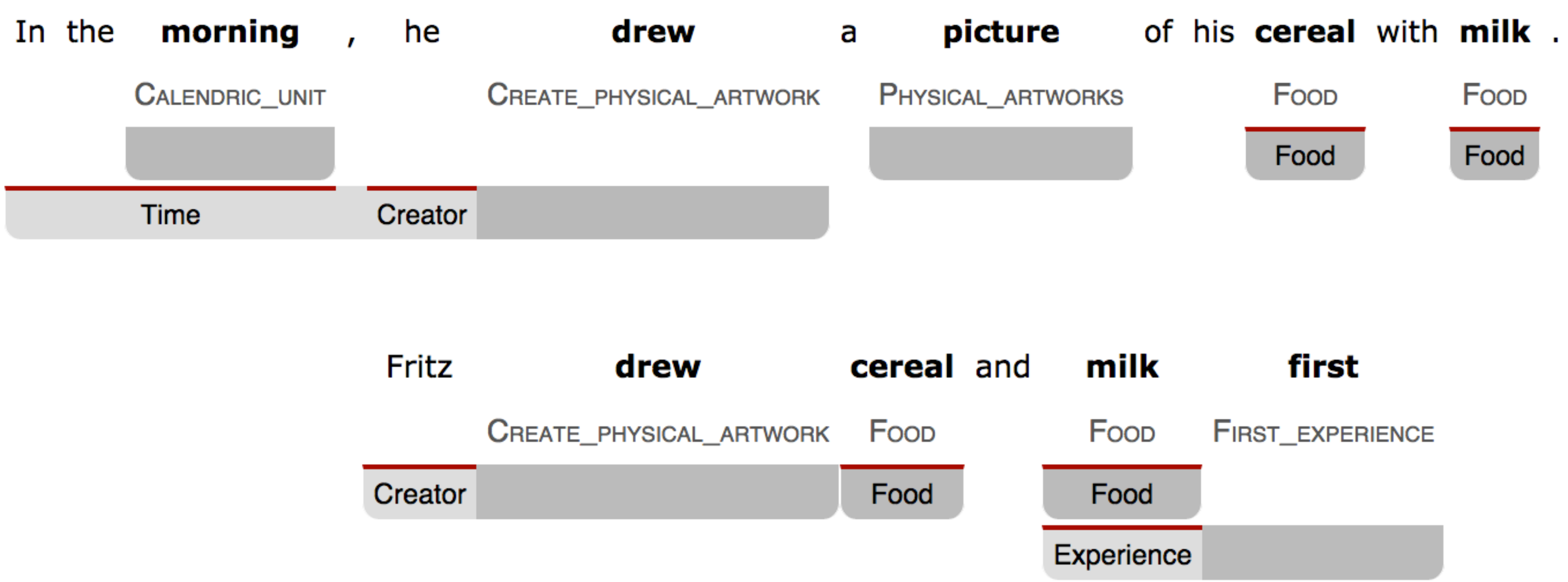
- dependency parsing
- **frame semantic parsing**



output of Carnegie Mellon frame semantic parser

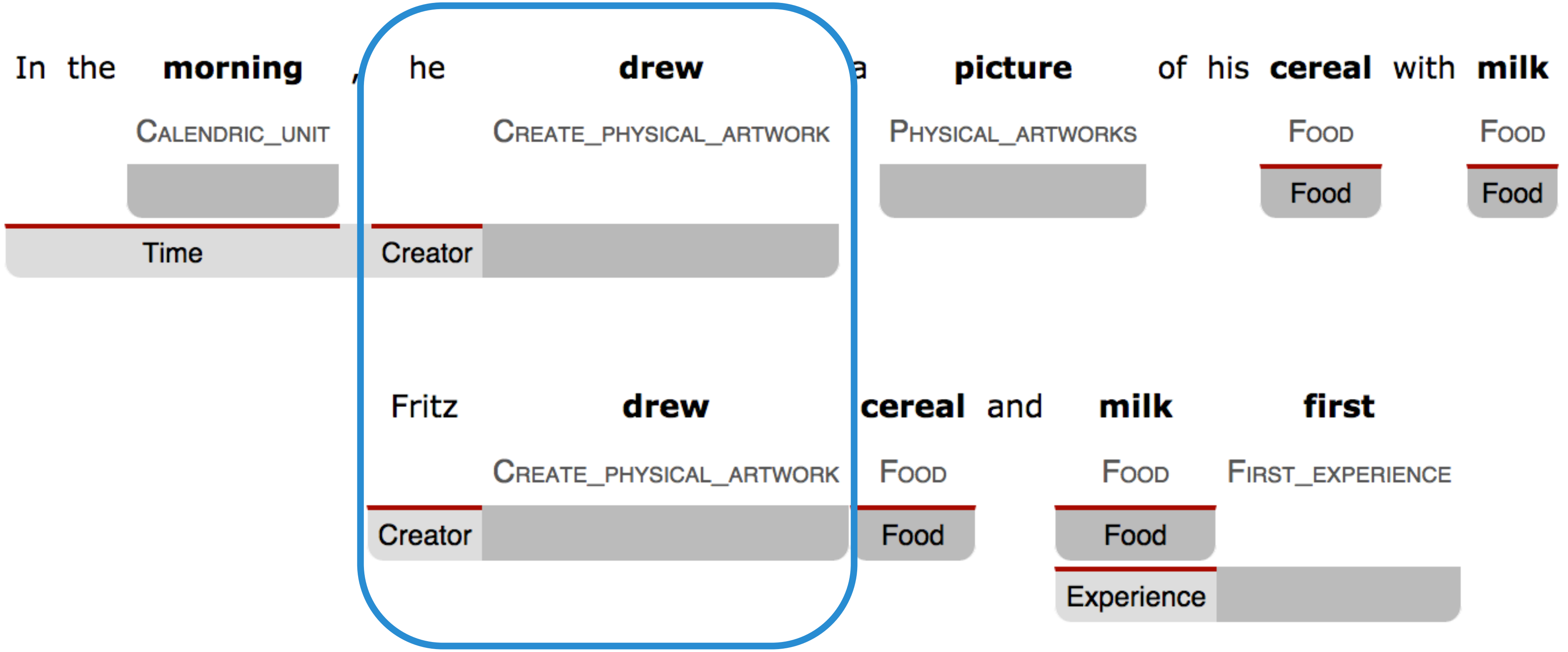
Our system uses several types of automatic linguistic analysis:

- dependency parsing
- **frame semantic parsing**



Our system uses several types of automatic linguistic analysis:

- dependency parsing
- **frame semantic parsing**

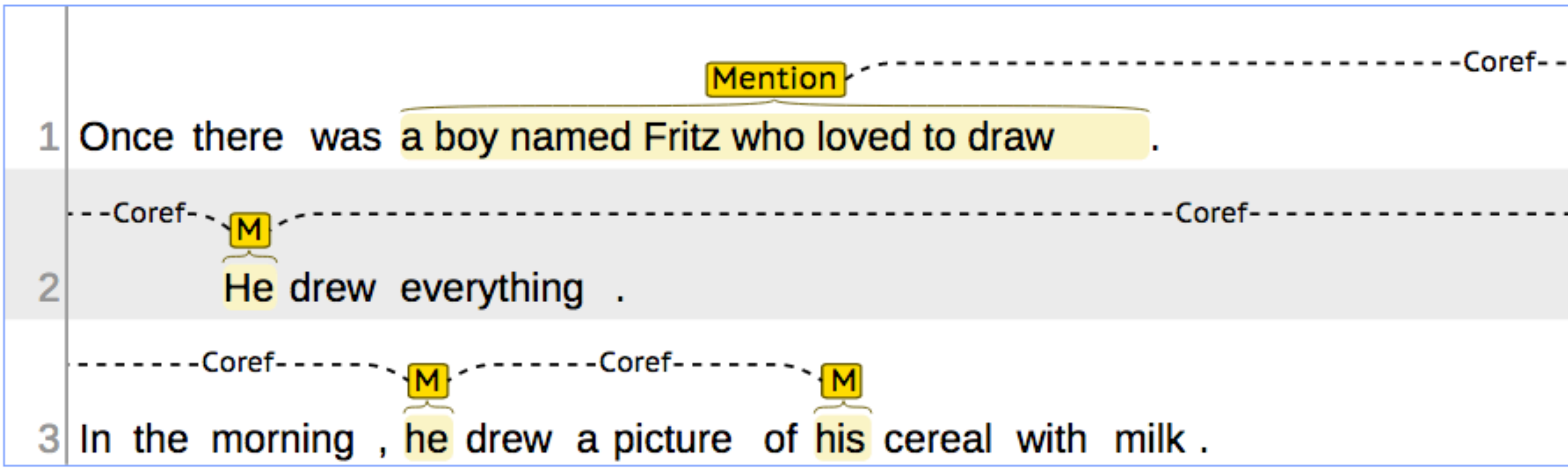


Our system uses several types of automatic linguistic analysis:

- dependency parsing
- frame semantic parsing
- **coreference**

Our system uses several types of automatic linguistic analysis:

– dependency parsing
– frame semantic parsing
– **coreference**



output of Stanford coreference resolution system

Our system uses several types of automatic linguistic analysis:

- dependency parsing
- frame semantic parsing
- coreference
- **word embeddings**

Once there was a boy named Fritz who loved to draw. He drew everything. In the morning, he drew a picture of his cereal with milk. His papa said, “Don’t draw your cereal. Eat it!”

…

What did Fritz draw first?

Once there was a boy named Fritz who loved to draw. He drew everything. In the morning, he drew a picture of his cereal with milk. His papa said, “Don’t draw your cereal. Eat it!”

…

What did Fritz draw first?

transformed question (using dependency parsing):

Fritz draw cereal and milk first

Fritz ≈ he (**coreference, frame semantics**)

draw ≈ drew (**word embeddings, frame semantics**)

with milk ≈ and milk (**word embeddings**)

Removing Features One at a Time