Evaluating MT output

Miloš Stanojević

Why do we need to evaluate MT output automatically?

- Rapid system development
- Tuning MT system
- Comparing different systems

Ideally we would use humans but they are too expensive.

So what is an evaluation metric?

- Basically a similarity function between the output of our system (“system translation”) and human translation (“reference translation(s)”)
- Similarity can be interpreted in different ways:
  - Overlap of sys and ref translation (precision, recall...)
  - Edit distance (insert, delete, move operations)
  - ...
- Different metrics make different choices on this matter

What kind of metric is the best?

- No consensus on that.
- Metrics make a good debating topic
- BLEU is de facto standard
  and everybody hates it
- Many alternatives but except METEOR and TER none gained popularity
- We will explain briefly BLEU and METEOR and then see everything from bigger picture
- In the end conclude with BEER (yet another metric)

BLEU

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

Plays the role of recall
Prevents too short translation

Corpus level
On sent level it's really bad
Smoothing needed

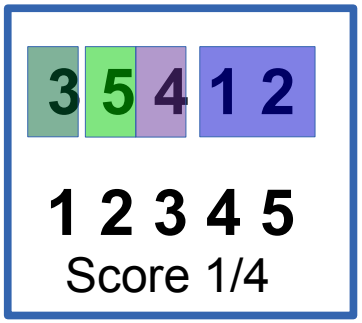
$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

What would happen if we had no brevity penalty? Any weird cases of translation that would be considered good but are actually bad?

What could be the problem with geometric mean?

Why use only precision explicitly and recall implicitly?

METEOR



Jump of an eye

$$Pen = \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}$$

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R}$$

$$Score = (1 - Pen) \cdot F_{mean}$$

Some tunable parameters estimated with hill climbing for correlation with humans

Additional resources (paraphrases, function words, word net, stemmers)

More linguistics needed?

- Characters
  - More robust BEER
- Words
  - We saw already few examples
- Syntax (dependency and constituency)
  - Dependency arcs matched, treelets matched...
- Semantics (semantic roles and paraphrases)
  - MEANT, SemPOS, Meteor (paraphrases), TERp

Question: Are higher levels of linguistic analysis necessarily better?

Weighting precision and recall

- All metrics have precision and recall in some way and they might weight them differently.
  - Ref: David Byrne is burning down the house.
  - Sys1: David Byrne is down the house.
  - Sys2: David Byrne is burning up and down the house.
- Do you prefer longer or shorter translation?
- Is that precision or recall?
- Imagine optimizing your system for precision/recall. What could get wrong?

The task

Human evaluator

Evaluation metric

Professional translator

MT system 1

MT system 2

MT system 3

Input sentence

The task

Human evaluator

Evaluation metric

Professional translator

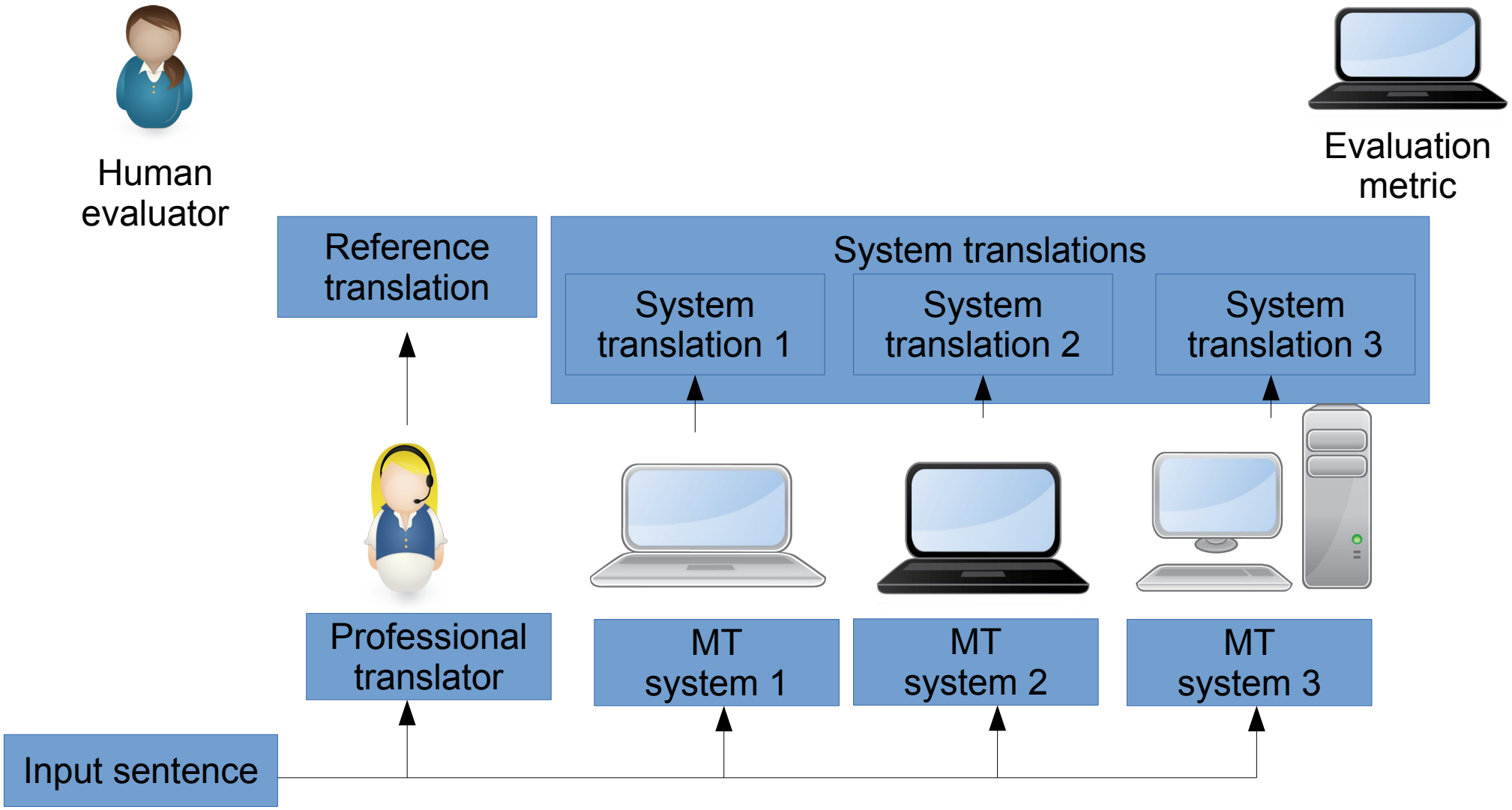
MT system 1

MT system 2

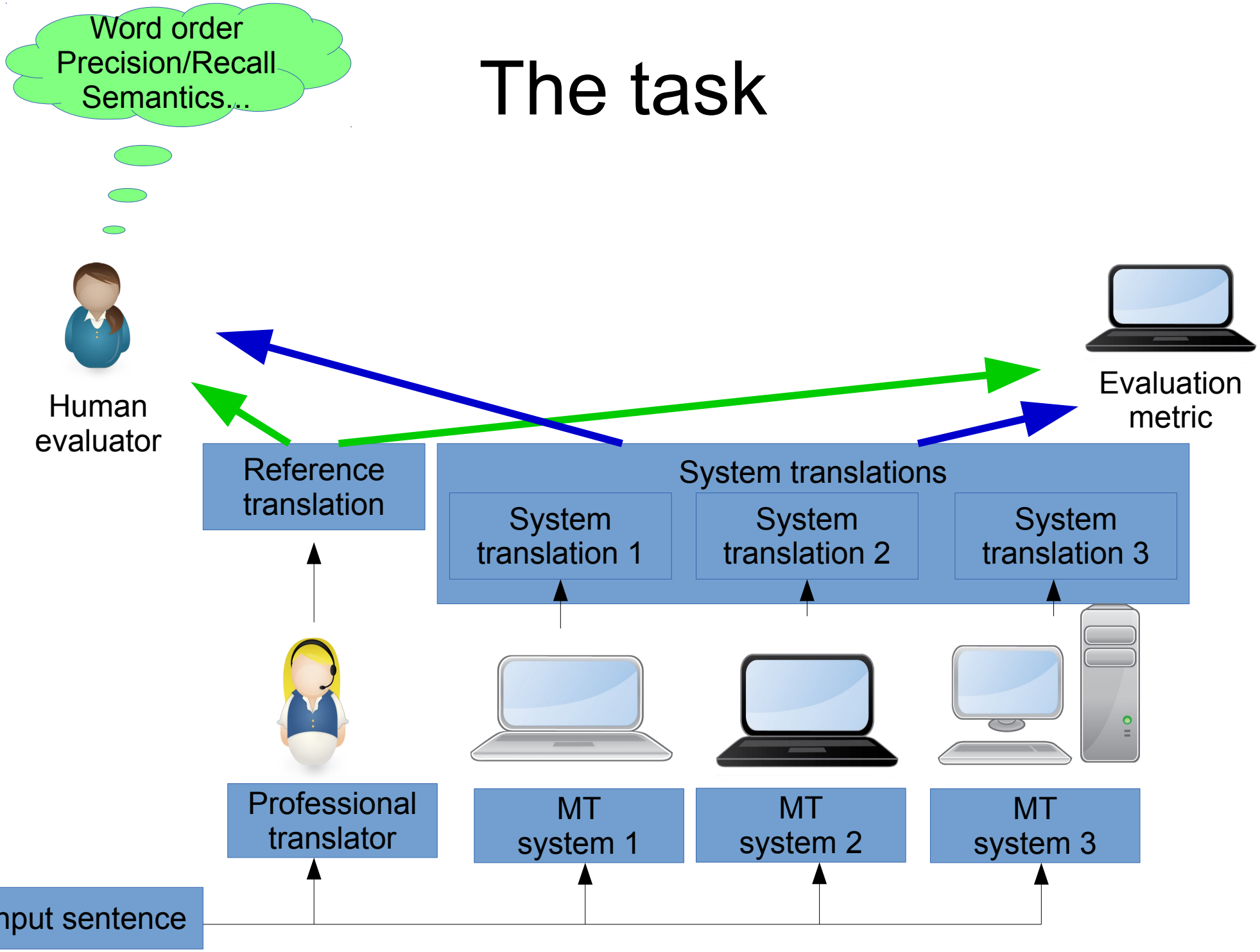
MT system 3

Input sentence

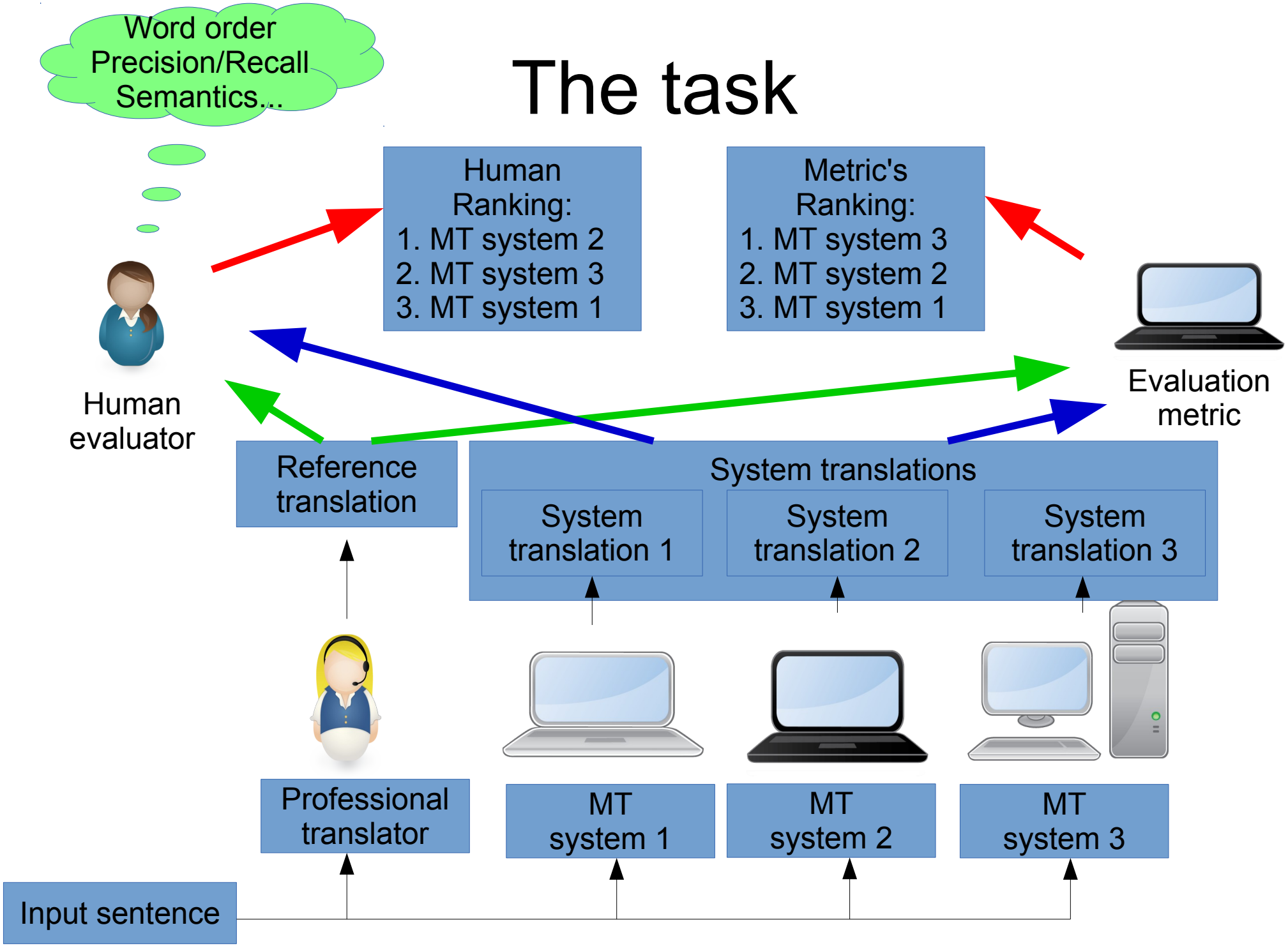
The task


Human evaluator
Evaluation metric
Reference translation
System translations
System translation 1
System translation 2
System translation 3
Professional translator
MT system 1
MT system 2
MT system 3
Input sentence

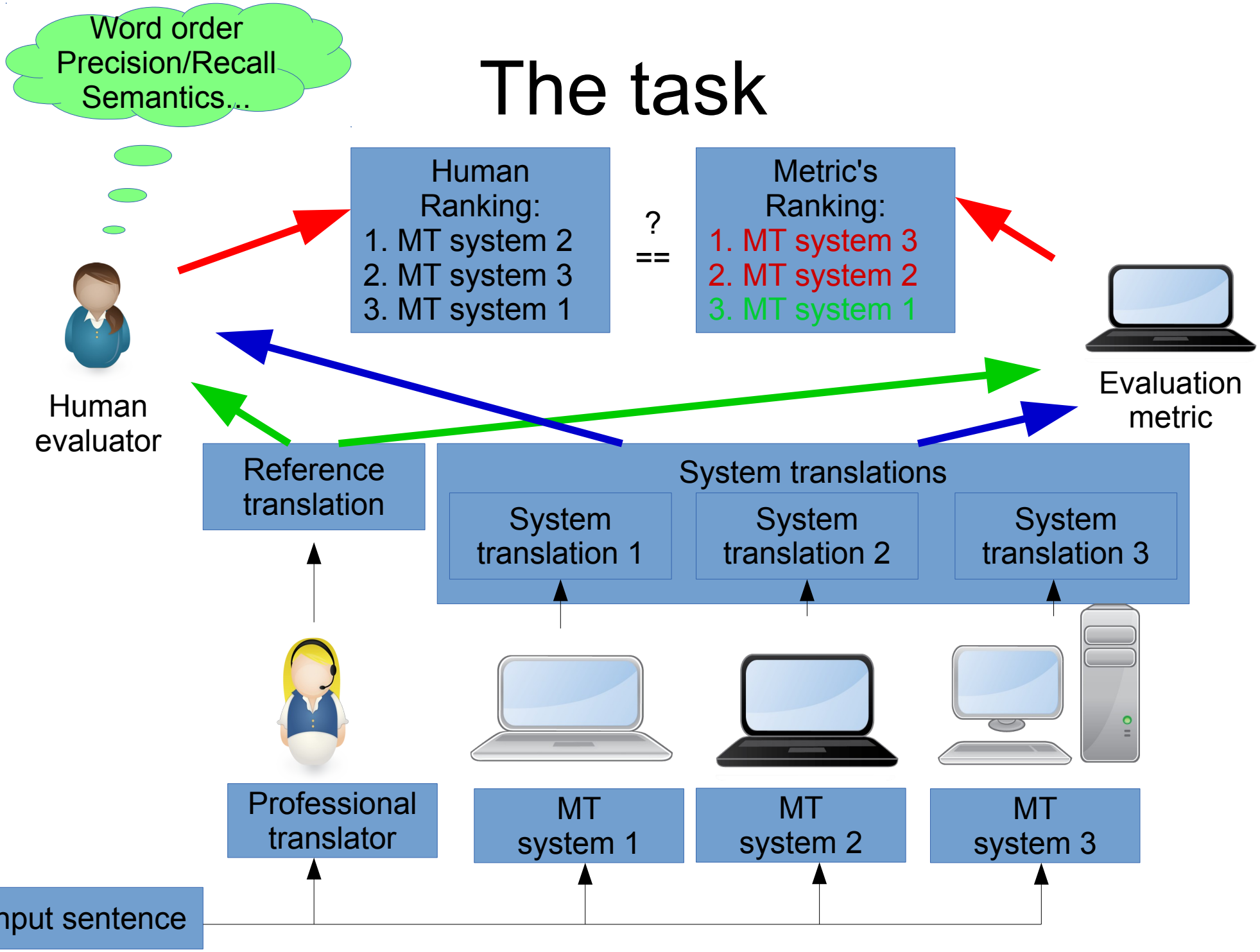
The task


Word order
Precision/Recall
Semantics...
Human evaluator
Evaluation metric
Reference translation
System translations
System translation 1
System translation 2
System translation 3
Professional translator
MT system 1
MT system 2
MT system 3
Input sentence

The task


Word order
Precision/Recall
Semantics...
Human Ranking:
1. MT system 2
2. MT system 3
3. MT system 1
Metric's Ranking:
1. MT system 3
2. MT system 2
3. MT system 1
Human evaluator
Evaluation metric
Reference translation
System translations
System translation 1
System translation 2
System translation 3
Professional translator
MT system 1
MT system 2
MT system 3
Input sentence

The task


Word order
Precision/Recall
Semantics...
Human Ranking:
1. MT system 2
2. MT system 3
3. MT system 1
?
==
Metric's Ranking:
1. MT system 3
2. MT system 2
3. MT system 1
Human evaluator
Evaluation metric
Reference translation
System translations
System translation 1
System translation 2
System translation 3
Professional translator
MT system 1
MT system 2
MT system 3
Input sentence

BEER
an example of a trained metric

Assume we use a linear model with features $\vec{\phi}$ and weight vector $\vec{\mathbf{w}}$. It would assign the score in the following way:

$$score(h, r) = \vec{\mathbf{w}} \cdot \vec{\phi}(h, r)$$

and that we have a human judgment that says that translation h_{good} is better than translation h_{bad}.

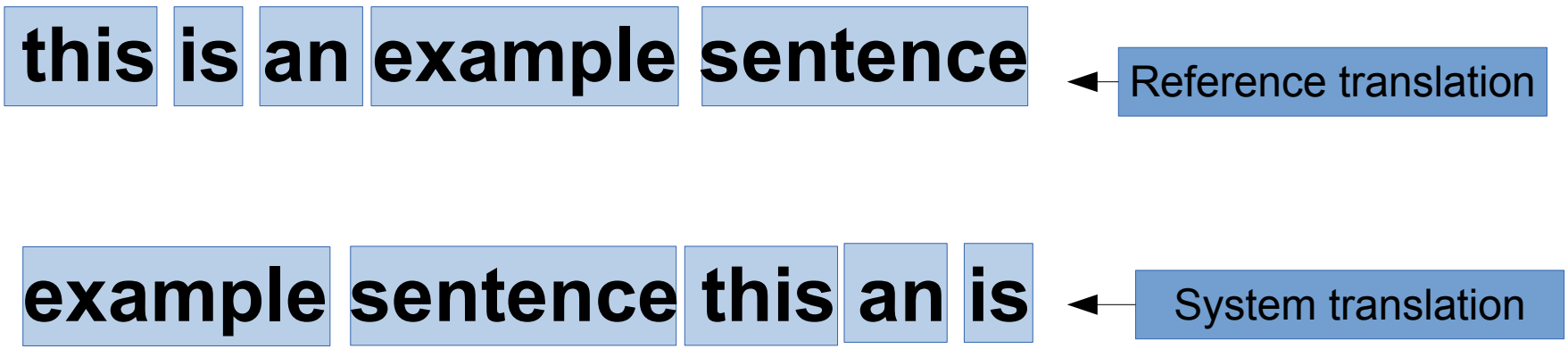
$$\begin{aligned} score(h_{good}, r) > score(h_{bad}, r) &\Leftrightarrow \\ \vec{\mathbf{w}} \cdot \vec{\phi}_{good} > \vec{\mathbf{w}} \cdot \vec{\phi}_{bad} &\Leftrightarrow \\ \vec{\mathbf{w}} \cdot \vec{\phi}_{good} - \vec{\mathbf{w}} \cdot \vec{\phi}_{bad} > 0 &\Leftrightarrow \\ \vec{\mathbf{w}} \cdot (\vec{\phi}_{good} - \vec{\phi}_{bad}) > 0 & \\ \vec{\mathbf{w}} \cdot (\vec{\phi}_{bad} - \vec{\phi}_{good}) < 0 & \end{aligned}$$

We transformed the ranking task into a standard classification task.

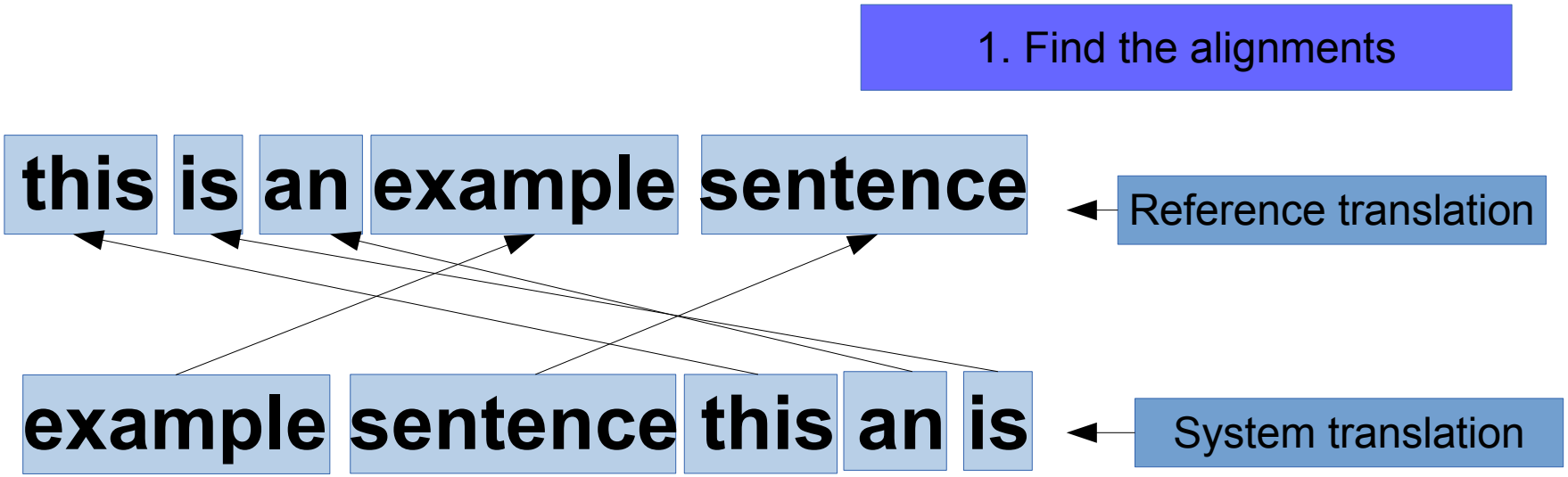
Lexical component

- Precision, recall, f-score on char n-grams
- Synonyms, paraphrases, function words...

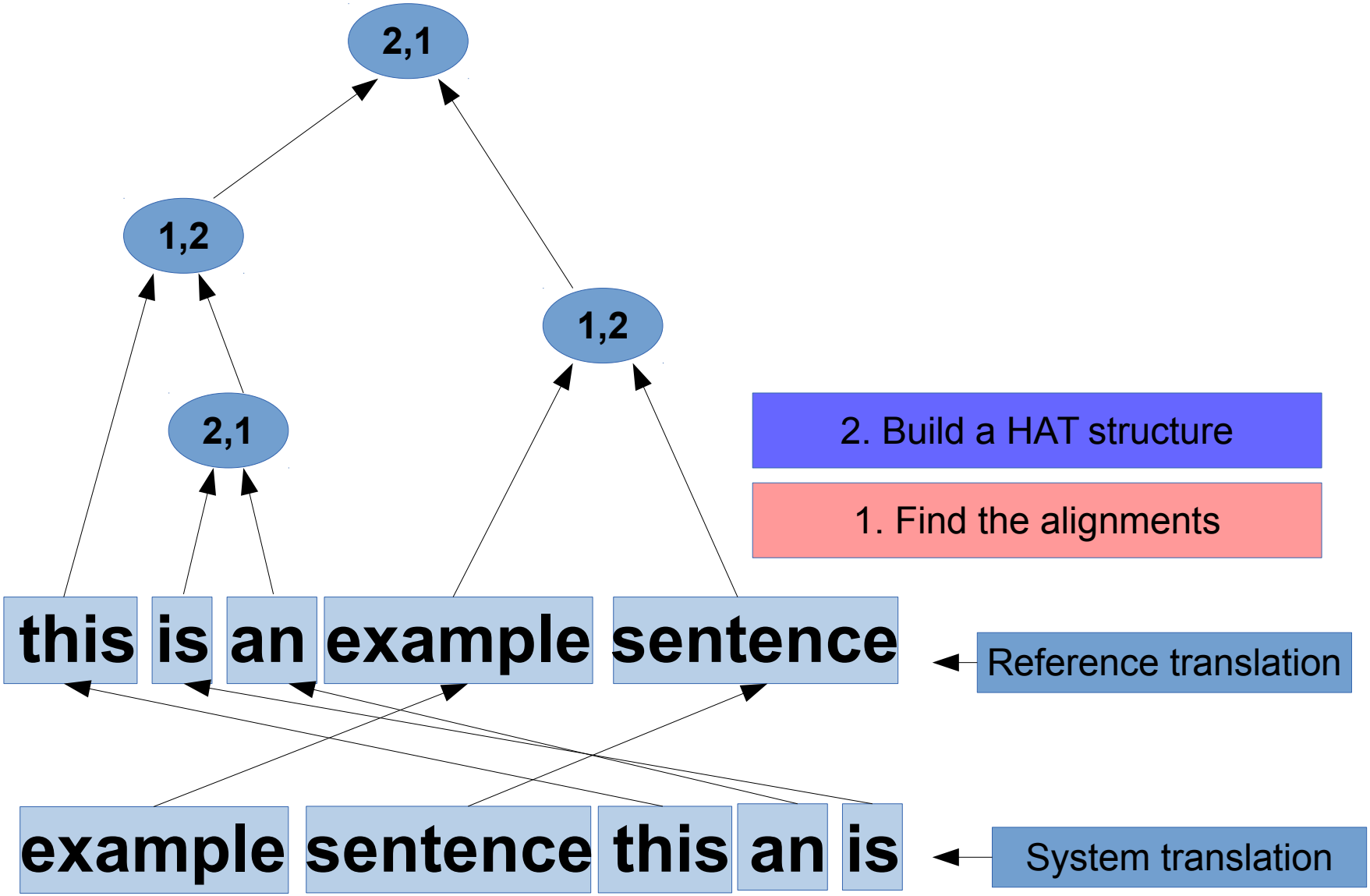
Reordering component



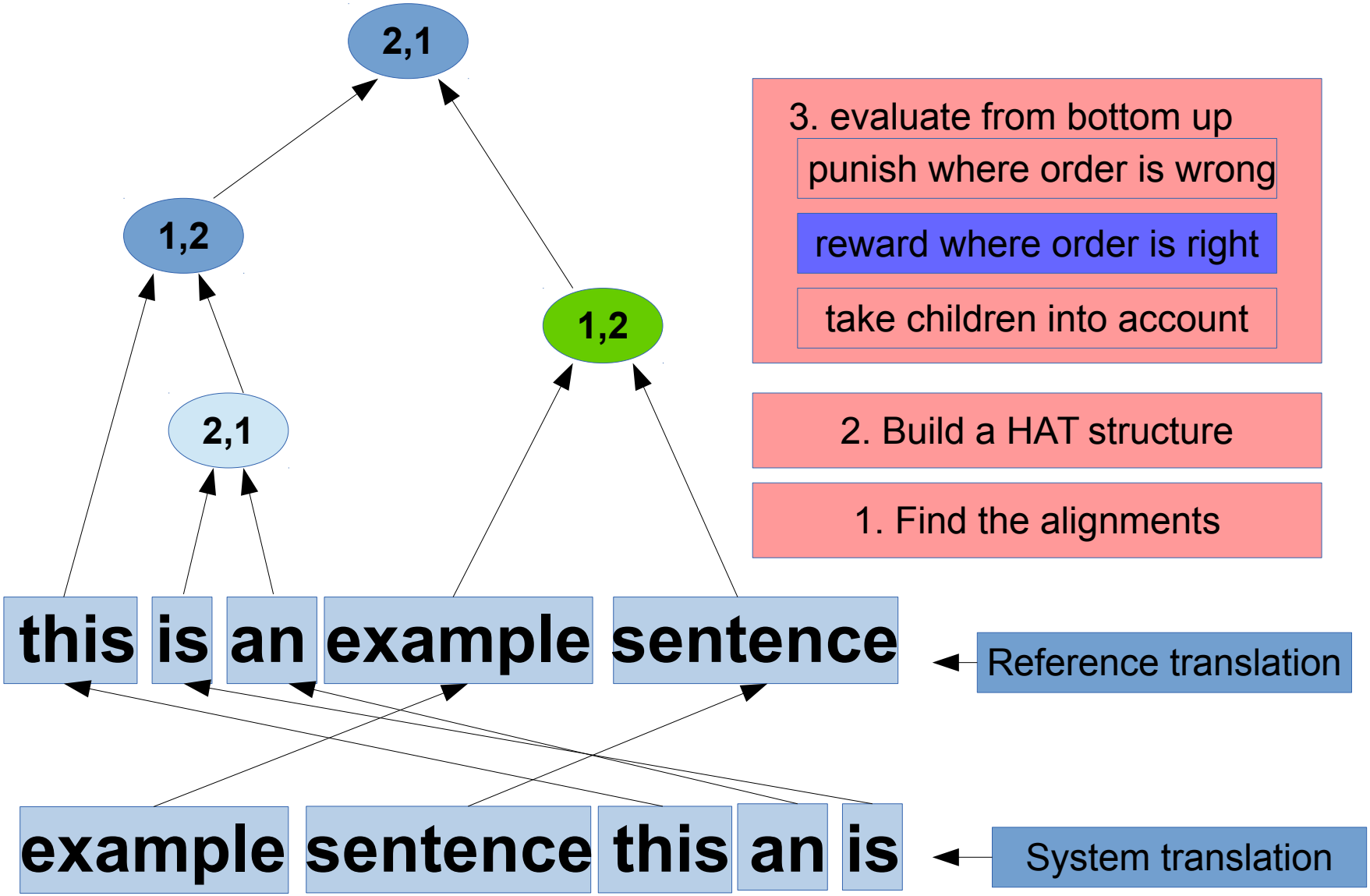
Reordering component



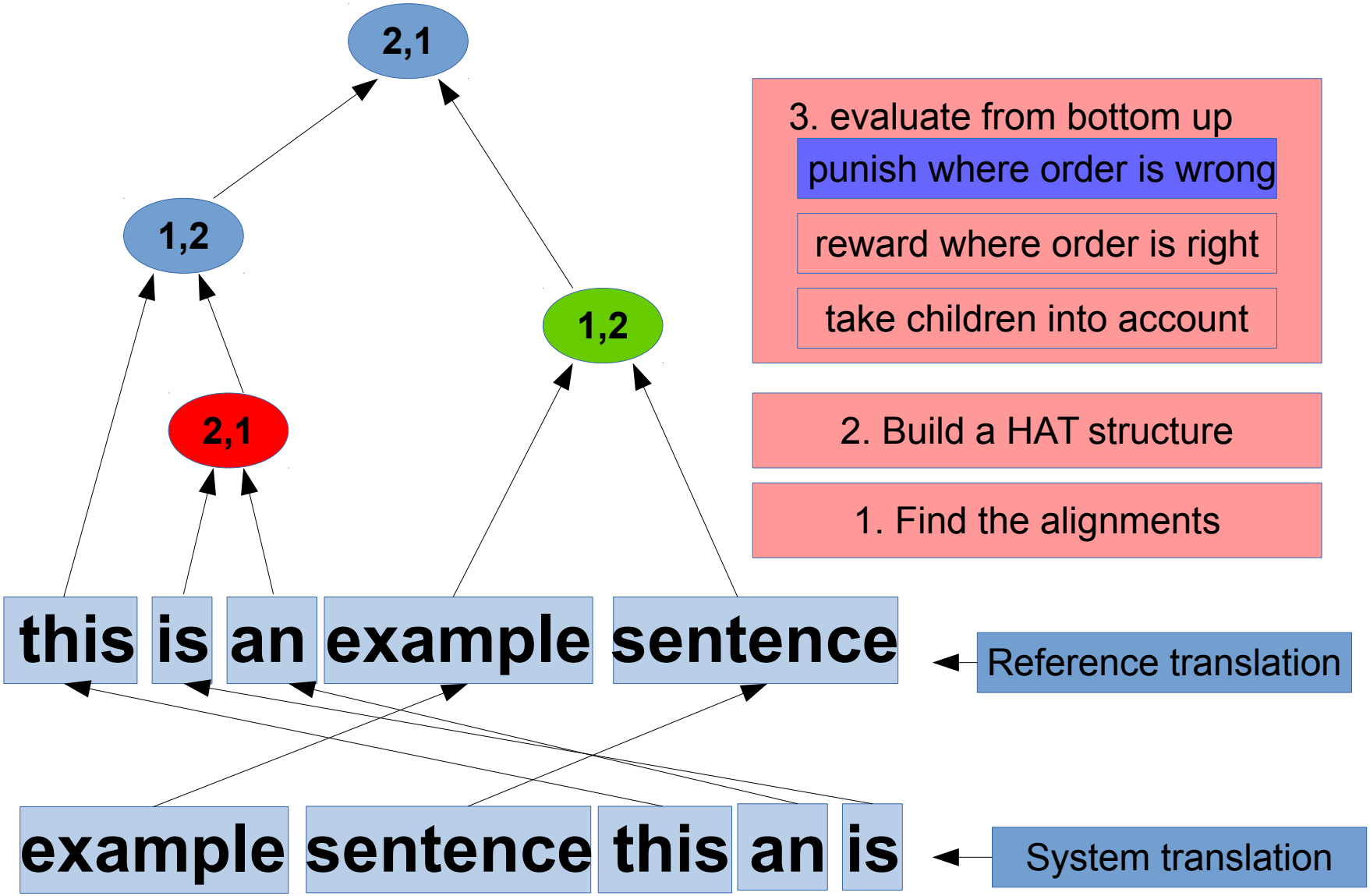
Reordering component



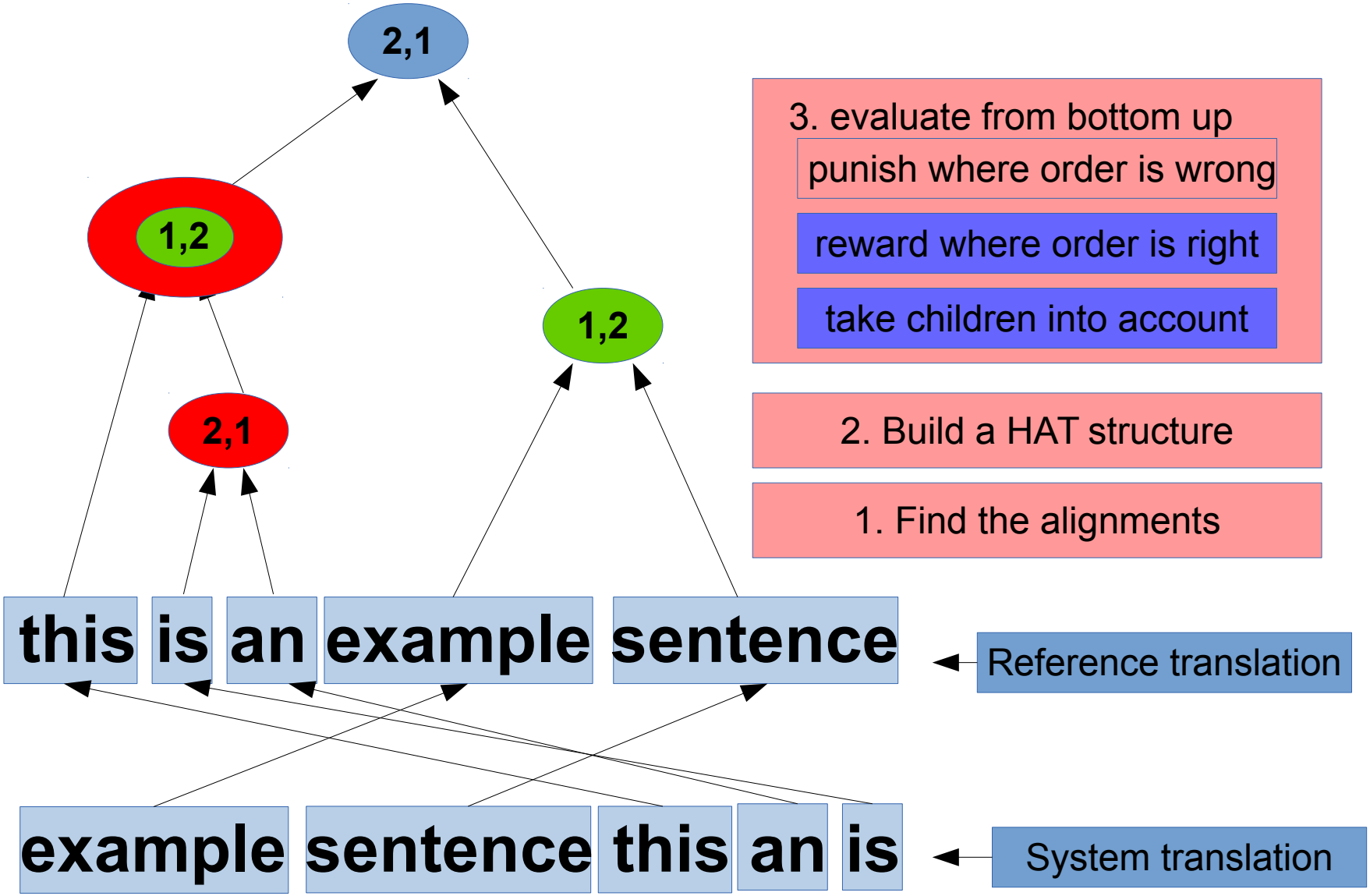
Reordering component



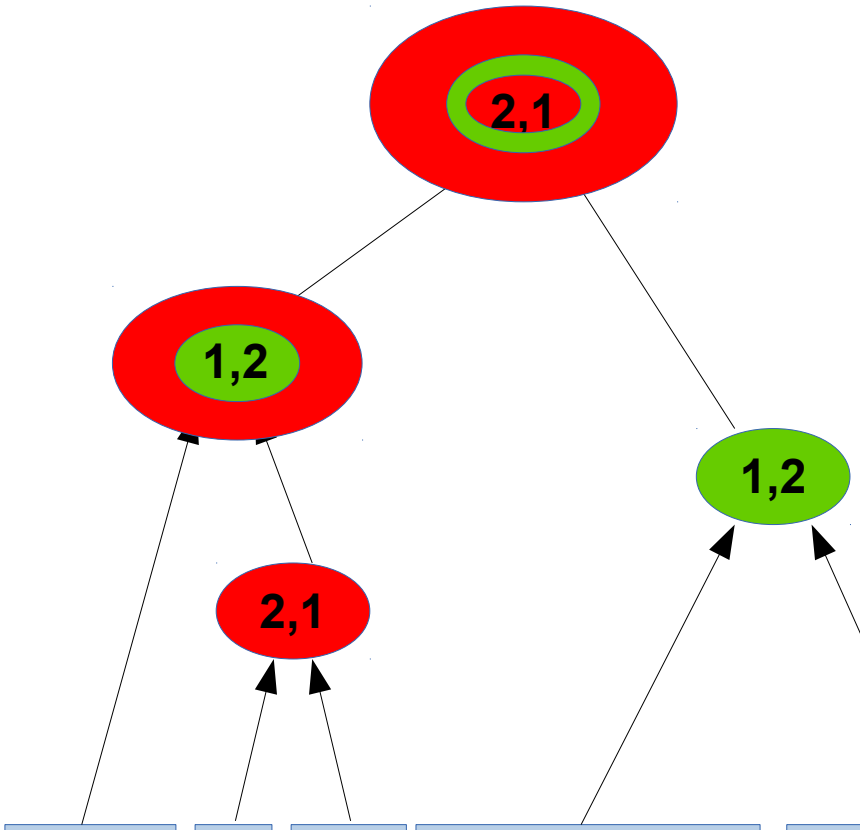
Reordering component



Reordering component



Reordering component



4. final score is
on the top of the HAT

3. evaluate from bottom up

punish where order is wrong

reward where order is right

take children into account

2. Build a HAT structure

1. Find the alignments

this is an example sentence ← Reference translation

example sentence this an is ← System translation

Some results

Direction Extracted-pairs	en-fr 25170	en-de 26760	en-hi 28120	en-cs 55880	en-ru 28960	Average	wmt12	wmt13	incl-human-ties
BEER	.295	.258	.250	.344	.440	.317	.313	.319	.270
METEOR	.278	.233	.264	.318	.427	.304	.281	.311	.271
AMBER	.261	.224	.286	.302	.397	.294	.268	.302	.265
BLEU-NRC	.257	.193	.234	.297	.391	.274	.233	.286	.253
APAC	.255	.201	.203	.292	.388	.268	.216	.283	.251
SENTBLEU-MOSES	.254	.185	.227	.290	.381	.268	.231	.278	.244
UPC-STOUT	.278	.224	n/a	.281	.425	.302	.298	.303	.253
UPC-IPA	.263	.217	n/a	.297	.426	.301	.289	.306	.257
REDSENT	.297	.236	n/a	n/a	n/a	.266	.246	.272	.255
REDCOMBSYSSENT	.290	.236	n/a	n/a	n/a	.263	.246	.268	.252
REDCOMBSENT	.290	.237	n/a	n/a	n/a	.263	.246	.268	.252
REDSYSSENT	.293	.229	n/a	n/a	n/a	.261	.232	.269	.252

Some results

Direction Extracted-pairs	fr-en 26090	de-en 25250	hi-en 20890	cs-en 21130	ru-en 24220	Average	wmt12	wmt13	incl-human-ties
DiscoTK-party-tuned	.433	.381	.434	.328	.364	.388	.388	.388	.304
BEER	.417	.337	.438	.284	.337	.363	.359	.364	.316
REDcombSent	.406	.338	.417	.284	.343	.357	.348	.361	.315
REDcombSysSent	.408	.338	.416	.282	.343	.357	.348	.361	.315
Meteor	.406	.334	.420	.282	.337	.356	.343	.360	.315
REDSysSent	.404	.338	.386	.283	.329	.348	.336	.352	.307
REDSent	.403	.336	.383	.283	.328	.347	.335	.351	.306
UPC-IPA	.412	.341	.367	.274	.324	.344	.341	.344	.298
UPC-STOUT	.403	.345	.351	.275	.324	.340	.338	.340	.292
VERTa	.399	.321	.386	.263	.318	.337	.321	.343	.302
VERTa-EQ	.407	.315	.384	.263	.313	.336	.323	.341	.299
DiscoTK-party	.395	.334	.362	.264	.313	.334	.334	.334	.261
AMBER	.367	.313	.362	.246	.296	.317	.302	.322	.283
BLEU-NRC	.382	.273	.322	.226	.273	.295	.267	.304	.270
sentBLEU-moses	.378	.271	.300	.213	.266	.286	.258	.294	.263
APAC	.364	.271	.288	.198	.276	.279	.243	.290	.259
DiscoTK-light	.311	.225	.237	.187	.212	.234	.234	.234	.183
DiscoTK-light-kool	.005	.001	.000	.002	.001	.002	-.996	.679	.221

Many other things to talk about

- Quality estimation (evaluation without references)
- Statistical testing
- Corpus vs. sentence level metrics
- Why metrics that are good for correlation with humans are not good for tuning?
- But we can talk about them some other time