idiap
INSTITUT DE RECHERCHE

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Using Coreference Links to Improve Spanish-to-English Machine Translation

Lesly Miculicich

Andrei Popescu-Belis

Content

1. Introduction
2. Coreference aware machine translation
3. Experiments and results
4. Conclusion

Content

1. Introduction
2. Coreference aware machine translation
3. Experiments and results
4. Conclusion

Motivation

Source:

*When she ran down, **the left slipper** remained stuck in the stairs, **it** was small and dainty.*

MT:

*Quand elle a couru, **la pantoufle gauche** est restée coincée dans les escaliers, **il** était petit et délicat.*

Motivation

Source: *Pertenezco a un **partido** político respetable.*

*– ¿Qué **partido**?*

Reference: *I belong to a respectable political **party**.*

*– Which **party**?*

MT: *I belong to a respectable political **party**.*

*– What a **match**?*

Machine Translation (MT)

$$\boldsymbol{e}_{best} = \underset{e}{argmax}\, p(\boldsymbol{e}|\boldsymbol{f})$$

Sentence in target language $\boldsymbol{e} = (e_1, e_2, \dots, e_n)$

Sentence in source language $\boldsymbol{f} = (f_1, f_2, \dots, f_m)$

Machine Translation (MT)

- Approaches:

  - **PBSMT**: Phase-based statistical machine translation
  - **NMT**: Neural machine translation

- Evaluation made comparing with <u>human translation as reference</u>. Common metric:

  - **BLEU**: *n*-gram precision

Coreference Resolution

- Linking or grouping mentions that refer to the same entity in a text.

  - **Mentions:** nouns, pronouns, noun-phrases, …
  - **Entities:** people, object, places, …
  - **Links**: coreference links, mention clusters, mention chains, …

- Evaluation made comparing with ground-truth. Common metrics:

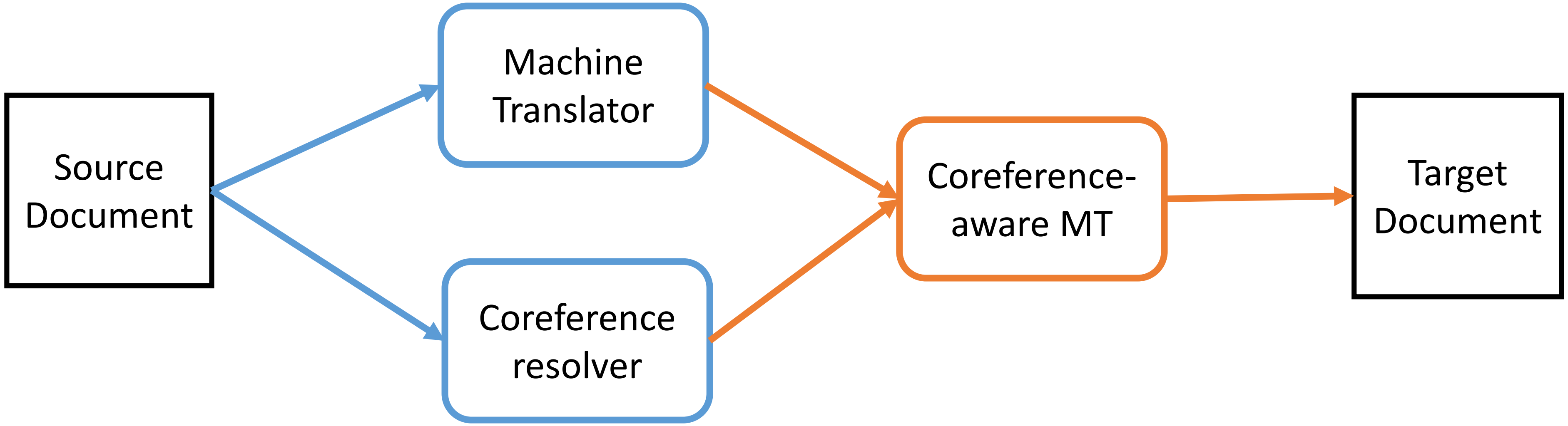
  - **MUC:** number of links to be inserted or deleted.
  - **B^3:** precision and recall at cluster-level for each mention.
  - **CEAF:** precision and recall at cluster-level for each entity.

Content

1. Introduction
2. Coreference aware machine translation
3. Experiments and results
4. Conclusion

Coreference-aware MT

- State-of-the-art
- Contribution



Objective: Improve the translation of documents by including coreference constraints.

Coreference in translation

Source (Spanish) [1]	Human Translation[2]	Machine Translation[2] [3]
La película narra la historia de [un joven parisiense]$_{c1}$ que marcha a Rumanía en busca de [una cantante zíngara]$_{c2}$, ya que [su]$_{c1}$ fallecido padre escuchaba siempre [sus]$_{c2}$ canciones. Pudiera considerarse un viaje fallido, porque [Ø]$_{c1}$ no encuentra [su]$_{c1}$ objetivo, pero el azar [le]$_{c1}$ conduce a una pequeña comunidad...	The film tells the story of [a young Parisian]$_{c1}$ who goes to Romania in search of [a gypsy singer]$_{c2}$, as [his]$_{c1}$ deceased father use to listen to [her]$_{c2}$ songs. It could be considered a failed journey, because [he]$_{c1}$ does not find [his]$_{c1}$ objective, but the fate leads [him]$_{c1}$ to a small community...	The film tells the story of [a young Parisian]$_{c1}$ who goes to Romania in search of [a gypsy singer]$_{c2}$, as [his]$_{c2}$ deceased father always listened to [his]$_{c2}$ songs. It could be considered [a failed trip]$_{c3}$ because [it]$_{c3}$ does not find [its]$_{c3}$ objective, but the chance leads to Ø a small community...

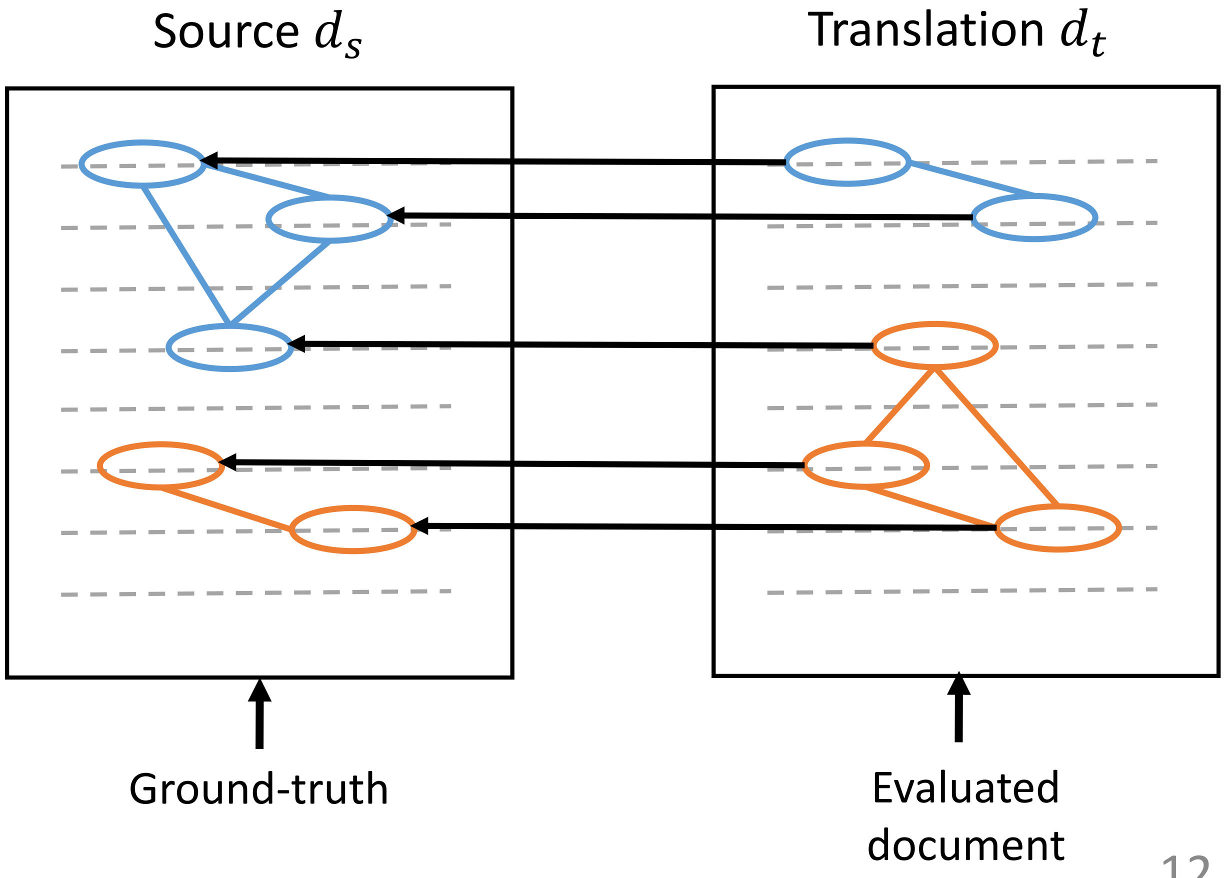
[1] Example from AnCora-CO with manual annotation of coreferences.

[2] Automatic coreference resolution with Stanford CoreNLP (http://stanfordnlp.github.io/CoreNLP/coref.html)

[3] Translation with a free online NMT

Defining Coreference Similarity Score

1. Apply coreference resolver on both sides.
2. Find alignments of mentions.
3. Calculate MUC, B3, and CEAF



Empirical Verification

Translation Quality

Coreference Quality

	BLEU	MUC	B^3	CEAF
Human translation	-	**37**	**32**	**41**
Commercial NMT	49.7	28	26	36
Baseline PBSMT	*43.4*	*23*	*24*	*33*

Values of F1 in %

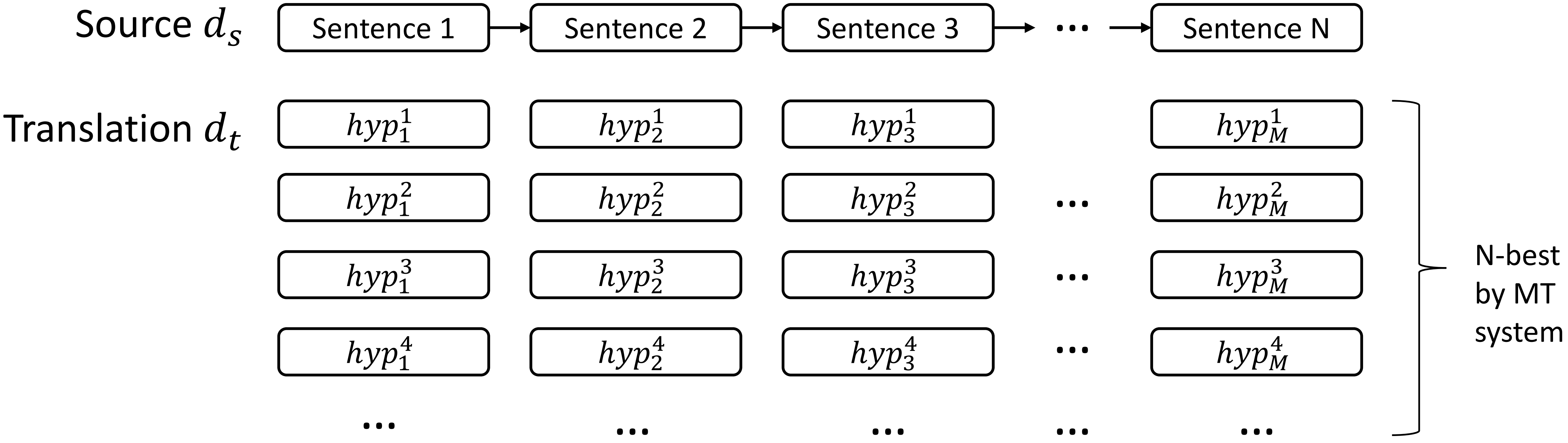
- Data: 3 K words from AnCora-CO with manual annotation of coreferences.
- Automatic coreference resolution with Stanford CoreNLP (http://stanfordnlp.github.io/CoreNLP/coref.html)
- Implementation of metrics from CoNLL 2012 (http://conll.cemantix.org/2012/)

Proposed approaches

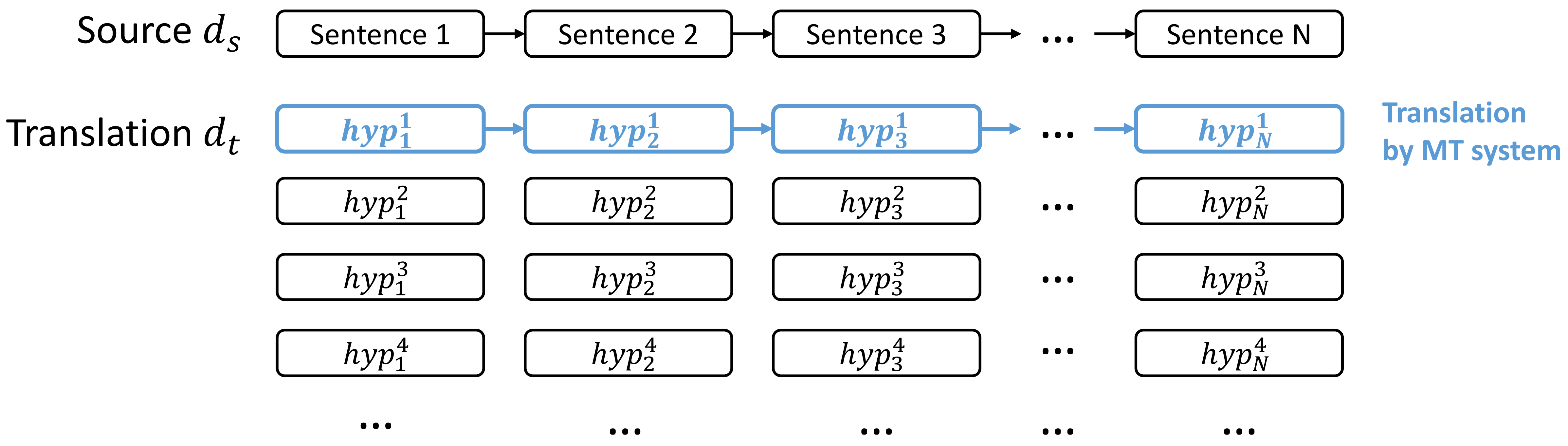
1. Re-ranking of *n*-best sentences
   → Changes at sentence-level
   → Scoring at document-level

2. Post-editing of mentions
   → Changes at mention-level
   → Scoring at cluster-level

Re-ranking


Source d_s
Sentence 1
Sentence 2
Sentence 3
…
Sentence N
Translation d_t
hyp_1^1 hyp_2^1 hyp_3^1 hyp_M^1
hyp_1^2 hyp_2^2 hyp_3^2 … hyp_M^2
hyp_1^3 hyp_2^3 hyp_3^3 … hyp_M^3
hyp_1^4 hyp_2^4 hyp_3^4 … hyp_M^4
… … … … …
N-best by MT system

Re-ranking



Source d_s
Sentence 1
Sentence 2
Sentence 3
...
Sentence N
Translation d_t
hyp_1^1
hyp_2^1
hyp_3^1
...
hyp_N^1
Translation by MT system
hyp_1^2
hyp_2^2
hyp_3^2
hyp_N^2
hyp_1^3
hyp_2^3
hyp_3^3
hyp_N^3
hyp_1^4
hyp_2^4
hyp_3^4
hyp_N^4

Re-ranking

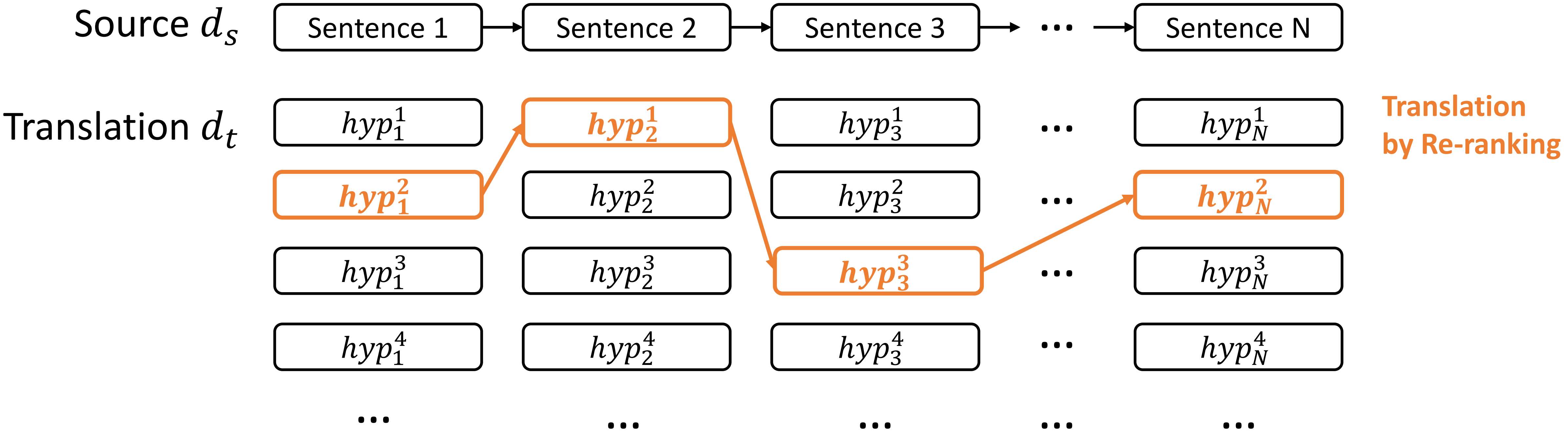
$$argmax\ C_{sim}(d_t, d_s) \qquad C_{sim} = (MUC + B^3 + CEAF)/3$$

Source d_s	Sentence 1 →	Sentence 2 →	Sentence 3 →	… →	Sentence M
Translation d_t	hyp_1^1	hyp_2^1	hyp_3^1	…	hyp_M^1
	hyp_1^2	hyp_2^2	hyp_3^2	…	hyp_M^2
	hyp_1^3	hyp_2^3	hyp_3^3	…	hyp_M^3
	hyp_1^4	hyp_2^4	hyp_3^4	…	hyp_M^4
	…	…	…	…	…

N-best
by MT
system

Re-ranking

$$argmax\ C_{sim}(d_t, d_s) \qquad C_{sim} = (MUC + B^3 + CEAF)/3$$



- ✓ Remove sentences with same set of mentions.
- ✓ Beam search

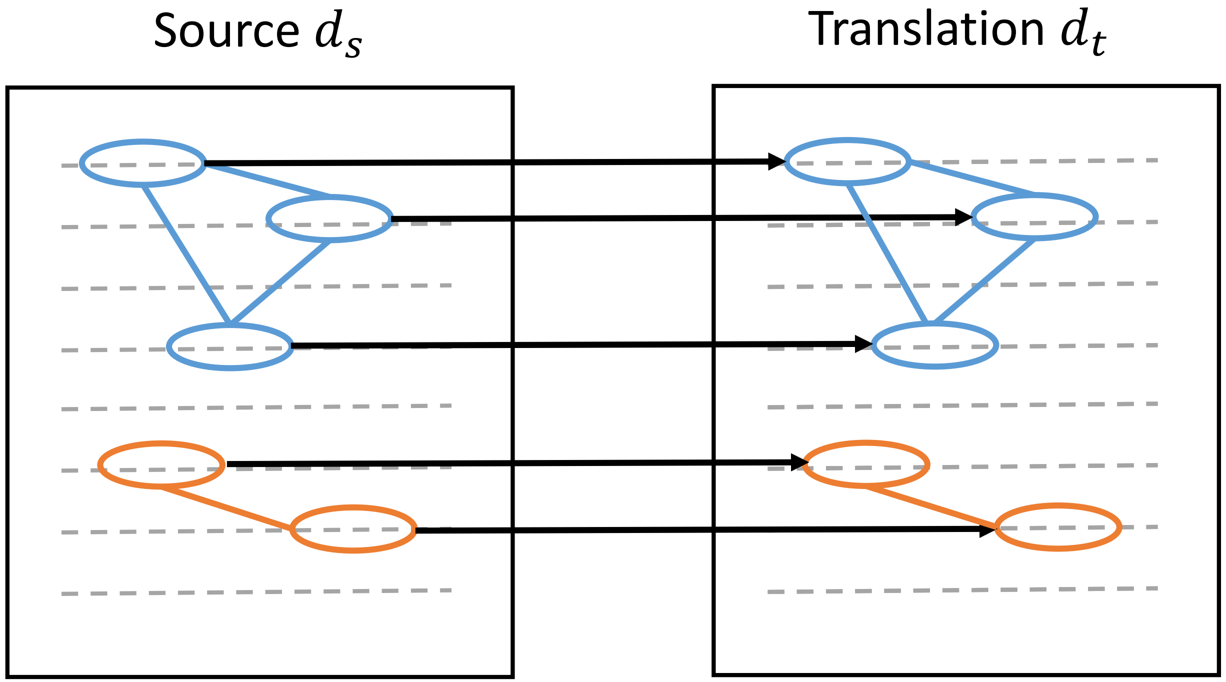
Re-ranking

✓Optimization at document-level.
✓Simple to use with a MT system.

✗ Not all mentions in a sentence can be optimized at the same time.
✗ Need to run coreference resolver at each step.

Post-editing

1. Apply coreference resolver on source side.
2. Find translation hypothesis of mentions in target side.
3. For each cluster: select the hypotheses that are more likely to refer to the same entity.



Post-editing

$argmax\ C_{score}(c_x)$

$C_{score}(c_x)$: Likelihood that all mentions in c_i refer to the same entity

Source cluster c_i	Mention 1	Mention 2	Mention 3	…	Mention M
Translation	hyp_1^1	hyp_2^1	hyp_3^1	…	hyp_M^1
	hyp_1^2	hyp_2^2	hyp_3^2	…	hyp_M^2
	hyp_1^3	hyp_2^3	hyp_3^3	…	hyp_M^3
	hyp_1^4	hyp_2^4	hyp_3^4	…	hyp_M^4
	…	…	…	…	…

N-best by MT system

Post-editing

Cluster score:

$$C_{score}(c_x) = C_s{}^{\lambda_1}.E_s{}^{\lambda_2}.T_s{}^{\lambda_3} \qquad \sum_i \lambda_i = 1$$

Elements in cluster

Entity representation from source

Translation frequency

Post-editing

Source cluster c_1	Partido politico	fue	partido	que
Translation	Political party	was	match	that
		It was	party	which
		He was		who
		She was		

N-best by MT system

Post-editing

Source cluster c_1	Partido politico	partido	que	fue
Translation	Political party	match	that	was
		party	which	It was
			who	He was
				She was

Reordering for number of options

Post-editing

$argmax\ C_{score}(c_x)$

$C_{score}(c_x)$: Likelihood that all mentions in c_i refer to the same entity

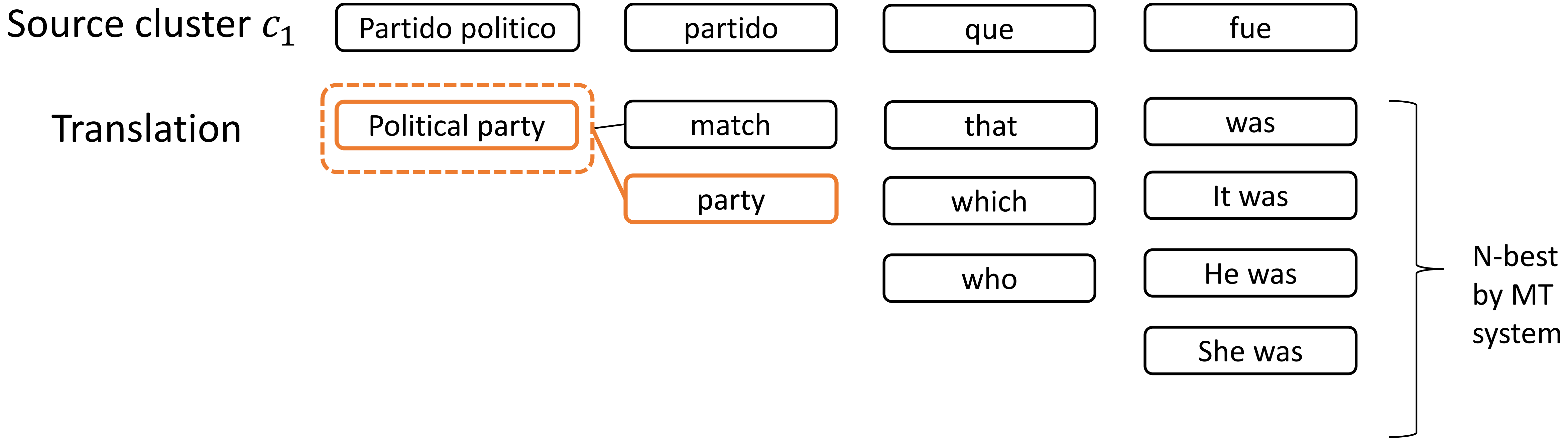
Source cluster c_1	Partido politico	partido	que	fue
Translation	Political party	match	that	was
		party	which	It was
			who	He was
				She was

N-best by MT system

Post-editing

$argmax\, C_{score}(c_x)$

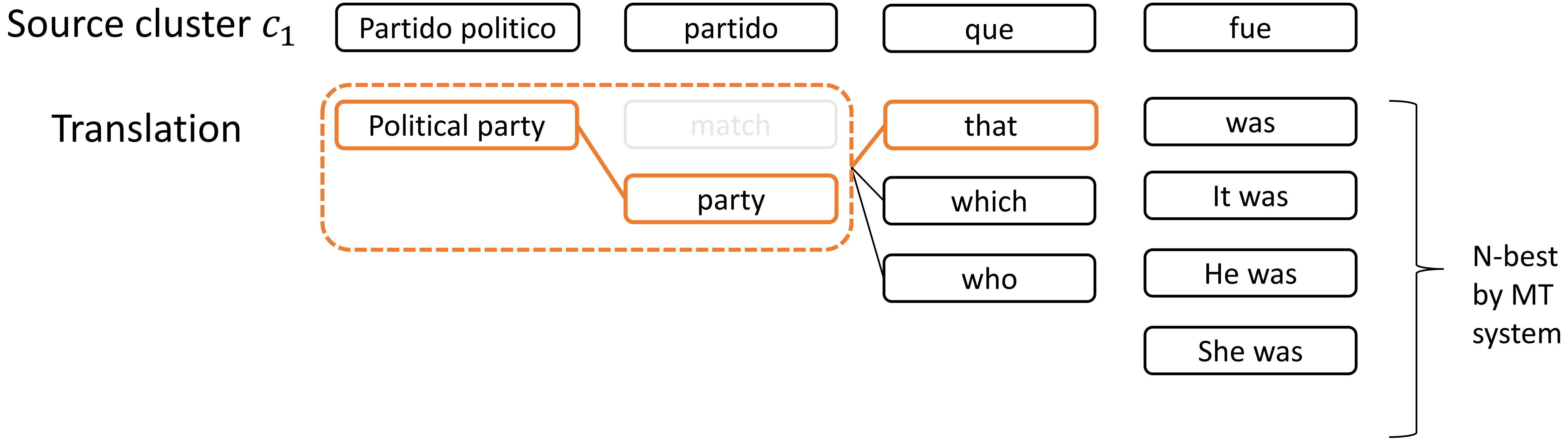
$C_{score}(c_x)$: Likelihood that all mentions in c_i refer to the same entity



Post-editing

$argmax\, C_{score}(c_x)$

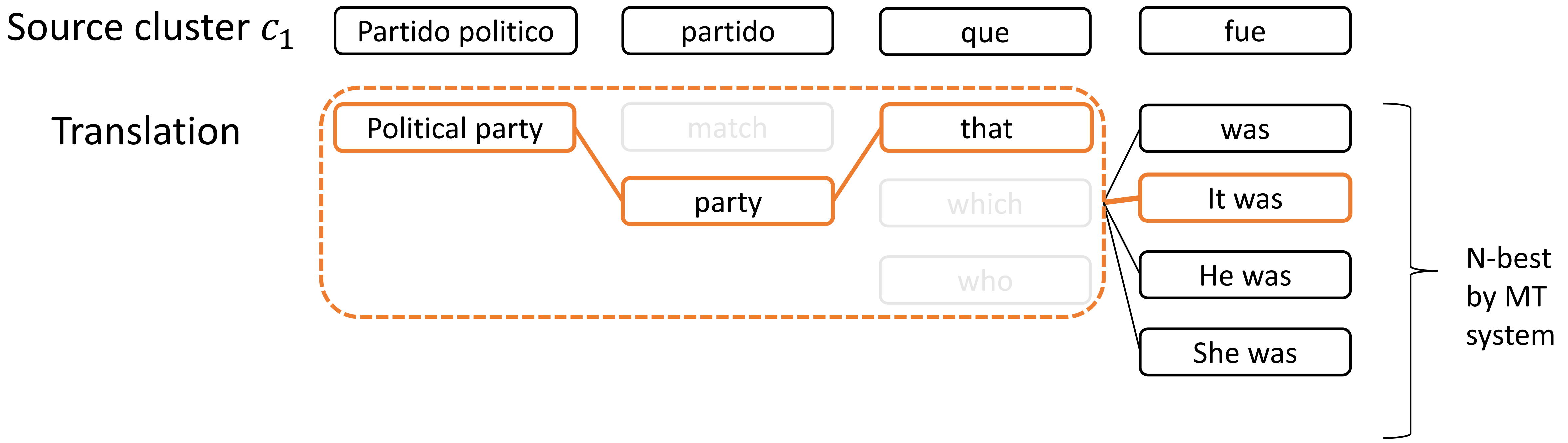
$C_{score}(c_x)$: Likelihood that all mentions in c_i refer to the same entity



Post-editing

$argmax\, C_{score}(c_x)$

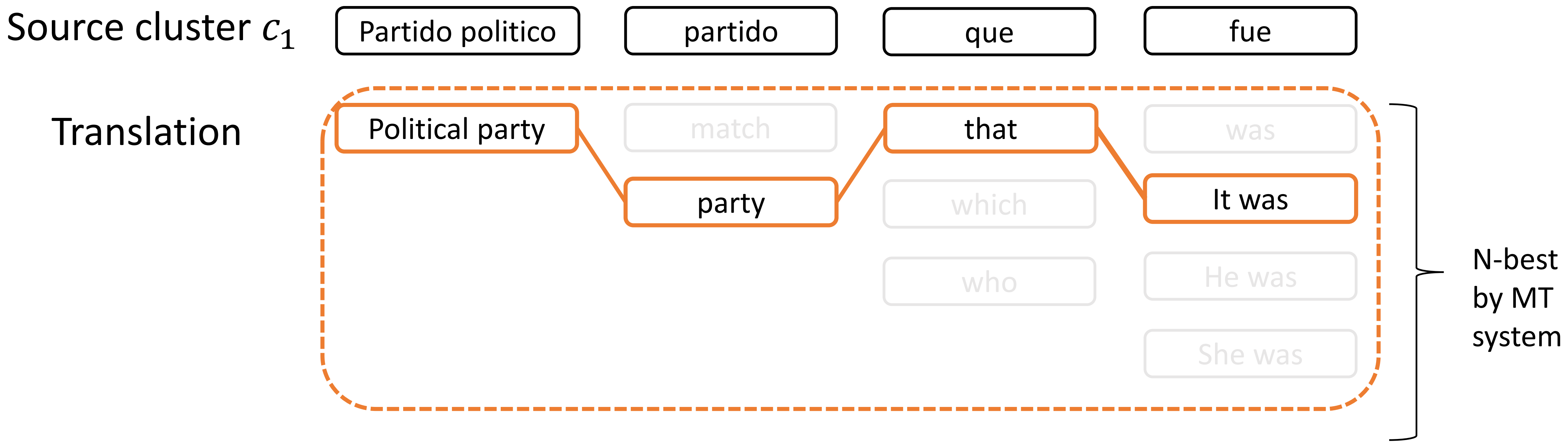
$C_{score}(c_x)$: Likelihood that all mentions in c_i refer to the same entity



Post-editing

$argmax\, C_{score}(c_x)$

$C_{score}(c_x)$: Likelihood that all mentions in c_i refer to the same entity



Content

1. Introduction
2. Coreference aware machine translation
3. Experiments and results
4. Conclusion

Baselines

System	Training[1]	Tuning[1 2]	Testing[1 3]	Language model	BLEU
$PBSMT_1$	1.9 M	5 K	3 K	3-gram 1.9 M	**24.51**
NMT_1	1.9 M	5 K	3 K	None	21.53
$PBSMT_2$	7.6 M	5 K	3 K	3-gram 7.6 M	25.43
NMT_2	7.6 M	5 K	3 K	None	**25.65**
$PBSMT_3$	14 M	5 K	3 K	4-gram 17 M	30.81
NMT_3	14 M	5 K	3 K	None	**32.21**

M: million sentences
K: thousand sentences

[1] Data from WMT 2013 Spanish-English.
[2] News-test 2010-2011
[3] News-test 2013

Evaluation Metrics

- BLEU

- APT: Accuracy of pronoun translation.

  Uses human translation as reference. It verifies:

  - Equal pronouns: exact match with reference.
  - Equivalent pronouns: learned from manual evaluation.

- ANT: Accuracy of noun translation

Evaluation

- State-of-the-art
- Contribution

Metric	PBSMT	NMT	PBSMT + Re-rank	PBSMT + Post-edit	PBSMT + Post-edit (automatic CR)
BLEU	46.5±4.3	46.9±3.7	41.7±3.9***	46.4±3.9	46.1±4.3
APT (pronouns)	0.35±0.07	0.37±0.07	**0.40±0.1***	**0.59±0.13***	**0.41±0.07***
ANT (nouns)	0.78±0.08	0.78±0.07	0.74±0.01***	0.78±0.07	0.76±0.09

Average and standard deviation over the test documents.
Statistical significance: * for 95.0%, ** for 99.0%, and *** for 99.9%

Human Evaluation

- State-of-the-art
- Contribution

Evaluation	PBSMT	PBSMT + Re-rank	PBSMT + Post-edit
Wrong	53	55	21
Acceptable	21	19	28
Identical to reference	115	115	140

Correctly Modified Example

Source:

[Barton]$_3$, por **[su]$_3$** parte , también dudó de la capacidad de [Megawati]$_2$ en **[su]$_2$** [nueva tarea] $_4$.

Reference:

[Barton] $_3$, for **[his] $_3$** part , also doubted [Megawati] $_2$ ’s ability in **[her] $_2$** [new task] $_4$.

Baseline:

[Barton] $_3$, for **[its] $_3$** part , also doubted the capacity of [Megawati] $_2$ in **[his] $_2$** [new task] $_4$.

Post-editing:

[Barton] $_3$, for **[his] $_3$** part , also doubted the capacity of [Megawati] $_2$ in **[her] $_2$** [new task] $_4$.

Correctly Modified Example

Source:

... que “ **[parece estar]$_2$** abrumada ... críticos consideran que **[no será]$_2$** capaz de hacerse con el papel de líder .

Reference:

...that “ **[she seems]$_2$** overwhelmed ... critics consider **[she will not be]$_2$** able to take the lead role .

Baseline:

... that “ **[appears to be]$_2$** overwhelmed ... critics believe that **[it will not be]$_2$** able to take a leading role . $_2$

Post-editing:

...that “ **[she seems]$_2$** to be overwhelmed ... critics believe that **[she will not be]$_2$** able to take a leading role

Content

1. Introduction
2. Coreference aware machine translation
3. Experiments and results
4. Conclusion

Conclusion

✓ Optimization at document-level including coreferences
✓ Post-editing approach improves pronouns translation

✗ Optimal solution (from reference) is not in the *n*-best hypothesis in ~20% of the cases
✗ Accuracy of coreference resolution is a limitation (~65% for English)

Future Work

- ✓ Testing on a larger dataset.
- ✓ Integration with the decoder of machine translation.
- ✓ Experiment application to neural machine translation.

Thanks