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)

\[
e_{\text{best}} = \arg\max_{e} p(e|f)
\]

Sentence in target language \( e = (e_1, e_2, ..., e_n) \)
Sentence in source language \( f = (f_1, f_2, ..., f_m) \)
Machine Translation (MT)

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

• Evaluation made comparing with human translation as reference. 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.
  • **B³:** 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

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

State-of-the-art

Contribution
## Coreference in translation

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

- Data: 3K 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/)

|                      | 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 %
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 $\rightarrow$ Sentence 2 $\rightarrow$ Sentence 3 $\rightarrow$ ... $\rightarrow$ Sentence N

Translation $d_t$

$hyp^1_1$ $\rightarrow$ $hyp^2_1$ $\rightarrow$ $hyp^3_1$ $\rightarrow$ $hyp^4_1$ $\rightarrow$ ... $\rightarrow$ $hyp^1_M$

$hyp^1_2$ $\rightarrow$ $hyp^2_2$ $\rightarrow$ $hyp^3_2$ $\rightarrow$ $hyp^4_2$ $\rightarrow$ ... $\rightarrow$ $hyp^2_M$

$hyp^1_3$ $\rightarrow$ $hyp^2_3$ $\rightarrow$ $hyp^3_3$ $\rightarrow$ $hyp^4_3$ $\rightarrow$ ... $\rightarrow$ $hyp^3_M$

$hyp^1_4$ $\rightarrow$ $hyp^2_4$ $\rightarrow$ $hyp^3_4$ $\rightarrow$ $hyp^4_4$ $\rightarrow$ ... $\rightarrow$ $hyp^4_M$

... $\rightarrow$ ... $\rightarrow$ ... $\rightarrow$ ... $\rightarrow$ ...

N-best by MT system
Re-ranking

Source $d_s$

Sentence 1 → Sentence 2 → Sentence 3 → ... → Sentence N

Translation $d_t$

Translation by MT system
Re-ranking

\[
\arg \max \ C_{sim}(d_t, d_s) \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad C_{sim} = (MUC + B^3 + CEAF)/3
\]

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

\[ \text{argmax } C_{\text{sim}}(d_t, d_s) \quad C_{\text{sim}} = \left( \frac{MUC + B^3 + CEAF}{3} \right) \]

Source \( d_s \):
- Sentence 1
- Sentence 2
- Sentence 3
- \( \ldots \)
- Sentence N

Translation \( d_t \):
- \( \text{hyp}_1^1 \)
- \( \text{hyp}_2^1 \)
- \( \text{hyp}_3^1 \)
- \( \ldots \)
- \( \text{hyp}_N^1 \)
- \( \text{hyp}_1^2 \)
- \( \text{hyp}_2^2 \)
- \( \text{hyp}_3^2 \)
- \( \ldots \)
- \( \text{hyp}_N^2 \)
- \( \text{hyp}_1^3 \)
- \( \text{hyp}_2^3 \)
- \( \text{hyp}_3^3 \)
- \( \ldots \)
- \( \text{hyp}_N^3 \)
- \( \text{hyp}_1^4 \)
- \( \text{hyp}_2^4 \)
- \( \text{hyp}_3^4 \)
- \( \ldots \)
- \( \text{hyp}_N^4 \)

- \( \ldots \)
- \( \ldots \)
- \( \ldots \)
- \( \ldots \)
- \( \ldots \)

- Translation by Re-ranking

✓ 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

\[ \text{argmax } C_{\text{score}}(c_x) \]

\[ C_{\text{score}}(c_x): \text{Likelihood that all mentions in } c_i \text{ refer to the same entity} \]

Source cluster \( c_i \):
- Mention 1
- Mention 2
- Mention 3
- \( \cdots \)
- Mention M

Translation:
- \( hyp_1^1 \)
- \( hyp_2^1 \)
- \( hyp_3^1 \)
- \( \cdots \)
- \( hyp_M^1 \)
- \( hyp_1^2 \)
- \( hyp_2^2 \)
- \( hyp_3^2 \)
- \( \cdots \)
- \( hyp_M^2 \)
- \( hyp_1^3 \)
- \( hyp_2^3 \)
- \( hyp_3^3 \)
- \( \cdots \)
- \( hyp_M^3 \)
- \( hyp_1^4 \)
- \( hyp_2^4 \)
- \( hyp_3^4 \)
- \( \cdots \)
- \( hyp_M^4 \)

\[ \text{N-best by MT system} \]
Post-editing

Cluster score:

\[ C_{Score}(c_x) = C_s^{\lambda_1} \cdot E_s^{\lambda_2} \cdot T_s^{\lambda_3} \]

\[ \sum_{i} \lambda_i = 1 \]

- 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 político
- 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

\[ \text{argmax } C_{\text{score}}(c_x) \]

\[ C_{\text{score}}(c_x): \text{Likelihood that all mentions in } c_i \text{ refer to the same entity} \]

Source cluster \( c_1 \)
- Partido politico
- partido
- que
- fue

Translation
- Political party
- match
- that
- was
- It was
- party
- which
- who
- He was
- She was

N-best by MT system
Post-editing

$$\arg \max \ C_{score}(c_x)$$

$C_{score}(c_x)$: 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
- It was
- which
- He was
- who
- She was

N-best by MT system
Post-editing

\[\text{argmax } C_{\text{score}}(c_x)\]

\[C_{\text{score}}(c_x)\]: Likelihood that all mentions in \(c_i\) refer to the same entity

Source cluster \(c_1\)
- Partido político
- partido
- que
- fue

Translation
- Political party
- match
- that
- was
- It was
- He was
- She was

N-best by MT system
Post-editing

\[ \text{argmax } C_{\text{score}}(c_x) \]

\[ C_{\text{score}}(c_x) \]: 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
- party: which
- that

N-best by MT system
- was
- It was
- He was
- She was
Post-editing

\[ \text{argmax} \ C_{\text{score}}(c_x) \]

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

Source cluster \( c_1 \)

Translation

Political party

\[ \text{match} \quad \text{that} \quad \text{It was} \]

party

\[ \text{which} \quad \text{He was} \]

\[ \text{who} \quad \text{She was} \]
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  |

\(^1\) Data from WMT 2013 Spanish-English.
\(^2\) News-test 2010-2011
\(^3\) News-test 2013

M: million sentences  
K: thousand sentences
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

| 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

| 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.

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