WMT 2016 Shared Task on Cross-lingual Pronoun Prediction

Liane Guillou, Christian Hardmeier, Preslav Nakov,
Sara Stymne, Jörg Tiedemann, Yannick Versley,
Mauro Cettolo, Bonnie Webber and Andrei Popescu-Belis

12/08/2016

Pronoun Translation Remains an Open Problem

- Pronoun systems do not map well between languages
  - E.g. grammatical gender for English $\rightarrow$ German
- Functional ambiguity:

anaphoric	I have an **umbrella**. **It** is red.
pleonastic	I have an umbrella. **It** is raining.
event	He lost his job. **It** came as a total surprise.

- SMT systems translate sentences in isolation
  - *Inter-sentential* anaphoric pronouns translated without knowledge of antecedent
- Two pronoun-related tasks at DiscoMT 2015:
  - Translation: systems failed to beat phrase-based baseline
  - Prediction: systems failed to beat language model baseline

Cross-Lingual Pronoun Prediction

- Given an input text and a translation with placeholders, replace the placeholders with pronouns
- Evaluated as a standard classification task

> Even though they were labeled whale meat , they were dolphin meat .
>
> Même si • avaient été étiquettés viande de baleine , • était de la viande de dauphin .
>
> 0-0 1-1 2-2 3-3 3-4 4-5 5-8 6-6 6-7 7-9 8-10 9-11 10-16 11-13 11-14 12-17
>
> Solution: *ils* *c'*

Task Overview

- DiscoMT 2015 English-French pronoun prediction task
  - Used fully inflected target-language text
- WMT 2016 tasks
  - Use lemmatised PoS-tagged target-language text
    Simulates SMT scenario in which we cannot trust inflection
- Four subtasks at WMT 2016:
  - English-French
  - French-English
  - English-German
  - German-English

Source and Target Pronouns

- Focus on source-language pronouns:
  - In **subject position**
  - That exhibit *functional ambiguity* ($\rightarrow$ multiple possible translations)

Source language	Pronouns
English	it, they
French	il, ils, elle, elles
German	er, sie, es

- **Prediction classes**: commonly aligned target-language translations

English-French Subtask: Pronouns

English subject pronouns	**French prediction classes**	
it	ce (inc. c')	[demonstrative]
they	cela (inc. ça)	[demonstrative]
	elle	[Fem. sg.]
	elles	[Fem. pl.]
	il	[Masc. sg.]
	ils	[Masc. pl.]
	on	[impersonal]
	OTHER	[anything else]

Data

- **Training data**:
  - News v9
  - Europarl v7
  - TED Talks (IWSLT 2015)
  - Automatic filtering of subject pronouns
- **Development data**: TED Talks
- **Test data**: TED Talks
  - Documents selected to ensure rare prediction classes are represented
  - Manual checks on subject pronoun filtering

```
elles  Elles  They arrive first .  REPLACE_0 arriver|VER en|PRP premier|NUM .|.  0-0
1-1 2-2 2-3 3-4
```

Figure : Example of training data format

Baseline System

- Baseline does what a typical SMT system would do:
  Predict everything with an n-gram model
- Fills REPLACE token gaps by using:
  - A fixed set of pronouns (prediction classes)
  - A fixed set of non-pronouns (OTHER words)
    Includes NONE (i.e., do not insert anything in the hypothesis)
- Configurable NONE penalty for empty slots to counterbalance the n-gram model's preference for brevity
- 5-gram language model provided for the task
- Similar language model baseline unbeaten at DiscoMT 2015

Evaluation

- **Macro-averaged Recall** - averaged over all classes to be predicted
  - DiscoMT 2015: Macro-averaged F-score
  - F-scores count each error twice
    once for precision; again for recall
- **Accuracy**
- Two official baseline scores provided for each subtask:
  - Default: NONE penalty set to zero
  - Optimised: NONE penalty tuned (for each subtask)

Submitted Systems

- 11 participants - some submitted to all subtasks
- Accepted primary and contrastive systems
- Two systems use LMs; all others use classifiers
- **Two main approaches**:
  - Use context from source and target text
    4 systems
  - Use source and target context + language-specific external tools / resources
    8 systems
- **Popular external tools**: coreference resolution, pleonastic "it" detection, dependency parsing

Results: English-French (Primary Systems)

	System	Macro-Avg Recall	Accuracy
1	**TurkuNLP**	**65.70_1**	**70.51_5**
2	**UU-Stymne**	**65.35_2**	**73.99_2**
3	**UKYOTO**	**62.44_3**	**70.51_4**
4	**uedin**	**61.62_4**	**71.31_3**
5	**UU-Hardmeier**	**60.63_5**	**74.53_1**
6	**limsi**	**59.32_6**	**68.36_7**
7	**UHELSINKI**	**57.50_7**	**68.90_6**
	baseline−1	*50.85*	*53.35*
8	**UUPPSALA**	**48.92_8**	**62.20_8**
	baseline0	*46.98*	*52.01*
9	**Idiap**	**36.36_9**	**51.21_9**

Results: English-German (Primary Systems)

	System	Macro-Avg Recall	Accuracy
1	**TurkuNLP**	**64.41** $_1$	**71.54** $_2$
2	**UKYOTO**	**52.50** $_2$	**71.28** $_3$
3	**UU-Stymne**	**52.12** $_3$	**70.76** $_4$
4	**UU-Hardmeier**	**50.36** $_4$	**74.67** $_1$
5	**uedin**	**48.72** $_5$	**66.32** $_6$
	baseline−2	*47.86*	*54.31*
6	**UUPPSALA**	**47.43** $_6$	**68.67** $_5$
7	**UHELSINKI**	**44.69** $_7$	**65.80** $_7$
8	**UU-Cap**	**41.61** $_8$	**63.71** $_8$
	baseline0	*38.53*	*50.13*
9	**CUNI**	**28.26** $_9$	**42.04** $_9$

Results: French-English (Primary Systems)

	System	Macro-Avg Recall	Accuracy
1	**TurkuNLP**	**72.03**$_1$	**80.79**$_2$
2	**UKYOTO**	**65.63**$_2$	**82.93**$_1$
3	**UHELSINKI**	**62.98**$_3$	**78.96**$_3$
4	**UUPSALA**	**62.65**$_4$	**74.39**$_4$
	baseline−1.5	*42.96*	*53.66*
	baseline0	*38.38*	*52.44*
5	**UU-Stymne**	**36.44**$_5$	**53.66**$_5$

Results: German-English (Primary Systems)

	System	Macro-Avg Recall	Accuracy
1	**TurkuNLP**	**73.91_1**	**75.36_3**
2	**UKYOTO**	**73.17_2**	**80.33_1**
3	**UHELSINKI**	**69.76_3**	**77.85_2**
4	**CUNI**	**60.42_4**	**64.18_6**
5	**UUPPSALA**	**59.56_5**	**73.71_4**
6	**UU-Stymne**	**59.28_6**	**69.98_5**
	baseline−1.5	*44.52*	*54.87*
	baseline0	*42.15*	*53.42*

Conclusions

- Most systems beat the baseline, in stark contrast with DiscoMT 2015
- En-Fr and En-De subtasks most popular
  - External tools / resources available for English
- RNNs work well for cross-lingual pronoun prediction
  - TurkuNLP: best system; all four subtasks
  - ukyoto: next best system; 3 subtasks
  - Systems use only source and target context
- uu-Stymne second place system for English-French

Next Steps

- **For Participants:**
  - Analyse and improve system performance
  - Integrate prediction systems into MT pipeline (post-editing, decoder feature, etc.)
- New task in 2017 [TBC]