Discourse in Machine Translation

Christian Hardmeier

Uppsala University

2018-09-04

Sentence-by-sentence Translation

It will go on sale in 2012.

Sentence-by-sentence Translation

It was presented last November and first demonstrated just a few days ago.
It will go on sale in 2012.

Sentence-by-sentence Translation

You can switch between them or have one for work and one for home. Both of them will run at the same time, says Srinivas Krishnamurti of VMware in an interview with Computer World magazine.

It was presented last November and first demonstrated just a few days ago.

It will go on sale in 2012.

Sentence-by-sentence Translation

This will let you have two user profiles at once on the same phone.
You can switch between them or have one for work and one for home. Both of them will run at the same time, says Srinivas Krishnamurti of VMware in an interview with Computer World magazine.
It was presented last November and first demonstrated just a few days ago.
It will go on sale in 2012.

Sentence-by-sentence Translation

This is the goal of the American company Vmware, which primarily develops computer virtualisation software.
This will let you have two user profiles at once on the same phone.
You can switch between them or have one for work and one for home. Both of them will run at the same time, says Srinivas Krishnamurti of VMware in an interview with Computer World magazine.
It was presented last November and first demonstrated just a few days ago.
It will go on sale in 2012.

Sentence-by-sentence Translation

Just press one key and in just a few seconds you can switch from Windows Mobile to Android.
This is the goal of the American company Vmware, which primarily develops computer virtualisation software.
This will let you have two user profiles at once on the same phone.
You can switch between them or have one for work and one for home. Both of them will run at the same time, says Srinivas Krishnamurti of VMware in an interview with Computer World magazine.
It was presented last November and first demonstrated just a few days ago.
It will go on sale in 2012.

Sentence-by-sentence Translation

Just press one key and in just a few seconds you can switch from Windows Mobile to Android.
This is the goal of the American company Vmware, which primarily develops computer virtualisation software.
This will let you have two user profiles at once on the same phone.
You can switch between them or have one for work and one for home. Both of them will run at the same time, says Srinivas Krishnamurti of VMware in an interview with Computer World magazine.
It was presented last November and first demonstrated just a few days ago.
It will go on sale in 2012.

Sentence-by-sentence Translation

Just press one key and in just a few seconds you can switch from Windows Mobile to Android.
This is the goal of the American company Vmware, which primarily develops computer virtualisation software.
This will let you have two user profiles at once on the same phone.
You can switch between them or have one for work and one for home. Both of them will run at the same time, says Srinivas Krishnamurti of VMware in an interview with Computer World magazine.
It was presented last November and first demonstrated just a few days ago.
It will go on sale in 2012.

Overview of this lecture

- Connectedness, cohesion and coherence
- Pronoun translation as an example: Challenges, evaluation and some results
- Cross-sentence modelling in NMT
- Final remarks

Texts are connected.

Discourse is more than a random set of utterances:
it shows connectedness.

(Sanders and Pander Maat, 2006)

Cohesion and Coherence

- **Cohesion:**
  reflected in overt linguistic elements and structures
- **Coherence:**
  reflected in the mental processes of the reader/listener

Types of Cohesion

1. Reference:
   John lives near the park. He often goes there.
2. Substitution:
   Dan loves strawberry ice-creams. He has one every day.
3. Ellipsis:
   All the children had an ice-cream today. Eva chose strawberry. Arthur had chocolate.
4. Conjunction:
   Eva walked into town, because she wanted an ice-cream.
5. Lexical cohesion:
   It was hot. Dan was lining up for an ice-cream.

Halliday and Hasan (1976). Examples from Sanders and Pander Maat (2006).

Coherence

- Reference, substitution, ellipsis and conjunction are identifiable linguistic processes, but lexical cohesion can be difficult to identify.
- Words in a discourse needn't be lexically similar to create a feeling of connectedness, but it must be possible to infer a plausible history from the text.

  Greenpeace has impeded a nuclear transportation in the Southern German state of Bayern. Demonstrators chained themselves to the rails.
- Connectedness can be seen as a characteristic of the mental representation of the text rather than the text itself: Coherence

What does it look like in practice?

- MT researchers have mostly focused on tangible problems of cohesion.
- **Reference:** Pronoun translation.
- **Conjunction:** Discourse connectives.
- **Lexical cohesion:** Domain adaptation. Vector space similarity.
- Verb tenses.
- General integration of “additional context”.

What does it look like in practice?

- MT researchers have mostly focused on tangible problems of cohesion.
- **Reference:** Pronoun translation.
- **Conjunction:** Discourse connectives.
- **Lexical cohesion:** Domain adaptation. Vector space similarity.
- Verb tenses.
- General integration of “additional context”.

Conjunction: Discourse connectives

Discourse connectives can have multiple senses
that may need disambiguation in translation.
(Meyer et al., 2012; Loáiciga, 2017)

The Champions League has become a source of income for clubs since it started in 1992.

since ≈ *because* or *since* ≈ *ever since*?

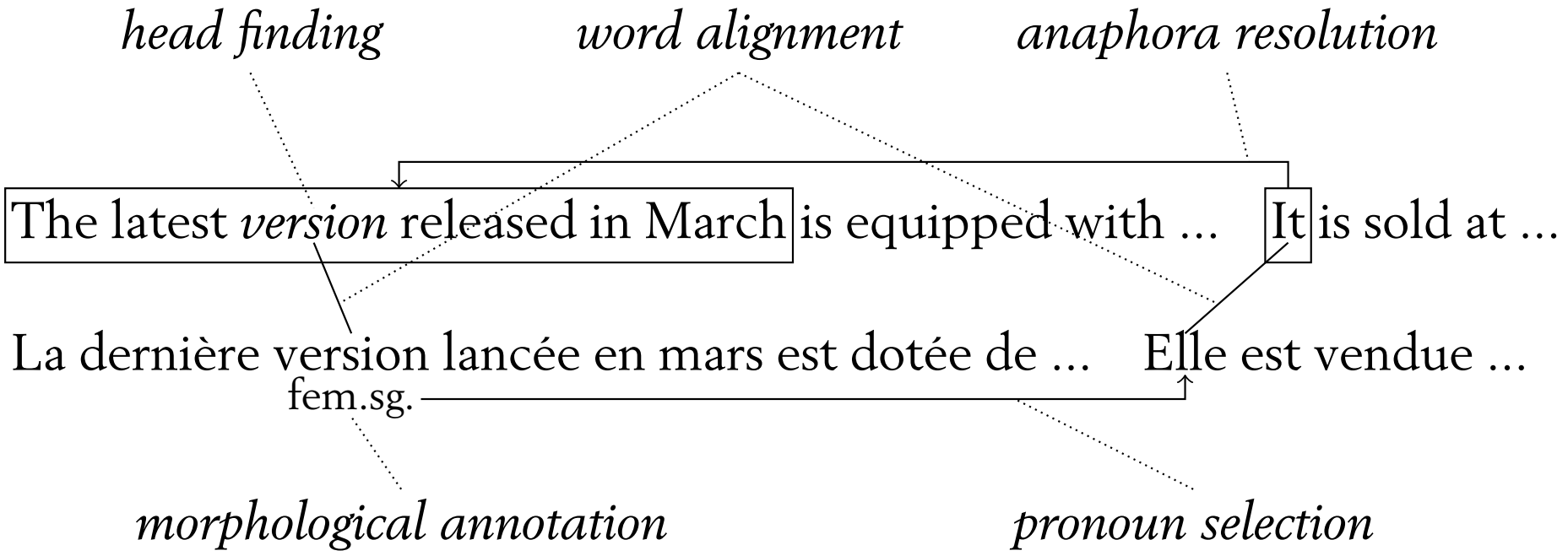
Lexical cohesion

- Much of the work on domain adaptation in MT can be seen as a way to improve lexical cohesion.
- People have tried to improve lexical cohesion by adding scores based on word similarity in vector spaces (Hardmeier, 2014; Martínez et al., 2017).
- Problem: It's not clear that MT actually has problems with consistency. Our goal must be to translate correctly, not consistently (Carpuat and Simard, 2012).
- Gender correctness is a similar problem (Vanmassenhove et al., 2017).

Pronoun translation

- Pronoun resolution was part of linguistic pipelines in the rule-based MT era.
- First papers on pronouns in SMT in 2010.
- Shared tasks on pronoun translation and prediction at DiscoMT 2015, WMT 2016, DiscoMT 2017.
- First, somewhat inconsistent signs of success in 2018.

Initial assumptions about pronoun translation



Target-side dependencies

The funeral of the Queen Mother will take place on Friday.
It will be broadcast live.

Target-side dependencies

The funeral of the Queen Mother will take place on Friday.
It will be broadcast live.

Les funérailles de la reine-mère auront lieu vendredi.
Elles seront retransmises en direct.

Target-side dependencies

The funeral of the Queen Mother will take place on Friday.
It will be broadcast live.

Les funérailles de la reine-mère auront lieu vendredi.
Elles seront retransmises en direct.

L'enterrement de la reine-mère aura lieu vendredi.
Il sera retransmis en direct.

Dangerous intuitions

- Common assumptions about pronouns:
  - Pronouns are linguistic elements that refer to something else in the text.
  - Pronouns agree with their antecedent.
  - Translating a pronoun requires generating a matching pronoun in the target language.
- All of this can be wrong.
- There are different types of pronouns.

Classification of pronouns

Classification of third person pronouns by Guillou (2016):

- **Anaphoric pronouns**
  *The infectious disease that's killed more humans than any other is malaria. **It**'s carried in the bites of infected mosquitos.*
- **Event pronouns**
  *But I think if we lost everyone with Down syndrome, **it** would be a catastrophic loss.*
- **Pleonastic pronouns**
  *And **it** seemed to me that there were three levels of acceptance that needed to take place.*

Properties of anaphoric pronominal references

- Types of pronouns
  - Personal
  - Demonstrative (proximal and distal)
  - Possessive
  - Reflexive
  - Relative
  - ...
- Grammatical function (subject or oblique)
- Intra-sentential vs. inter-sentential reference

Special cases

- **Singular they:** used to refer to a single person without specifying their gender.
- **Collective nouns:** Some entities can be conceptualised as either singular or plural.

  The company wanted its/their money back.

General-purpose MT metrics and pronouns

- Metrics like BLEU may not be sensitive to pronoun translation.
- They cannot keep track of target-side dependencies.
- They are totally unspecific.

Source:	*Translation:*
Until the 1980s , the farm was in the hands of the Argentinians .	Jusque dans les années 80 , la ferme est entre les mains des Argentins .
They raised beef cattle on what was essentially wetlands .	Ils ont soulevé des bovins de boucherie sur ce qui était essentiellement des zones humides .
They did it by draining the land .	Ils l' ont fait par l' assèchement des terres .
They built this intricate series of canals , and they pushed water off the land and out into the river .	Ils ont construit cette série complexe de canaux , et ils ont poussé l' eau du sol et dans la rivière .
Well , they couldn 't make it work , not economically .	Eh bien , ils ne pouvaient pas le faire fonctionner , pas économiquement .
And ecologically , **it** was a disaster .	Et sur le plan écologique , **XXX** fut un désastre .

Select the correct pronoun:

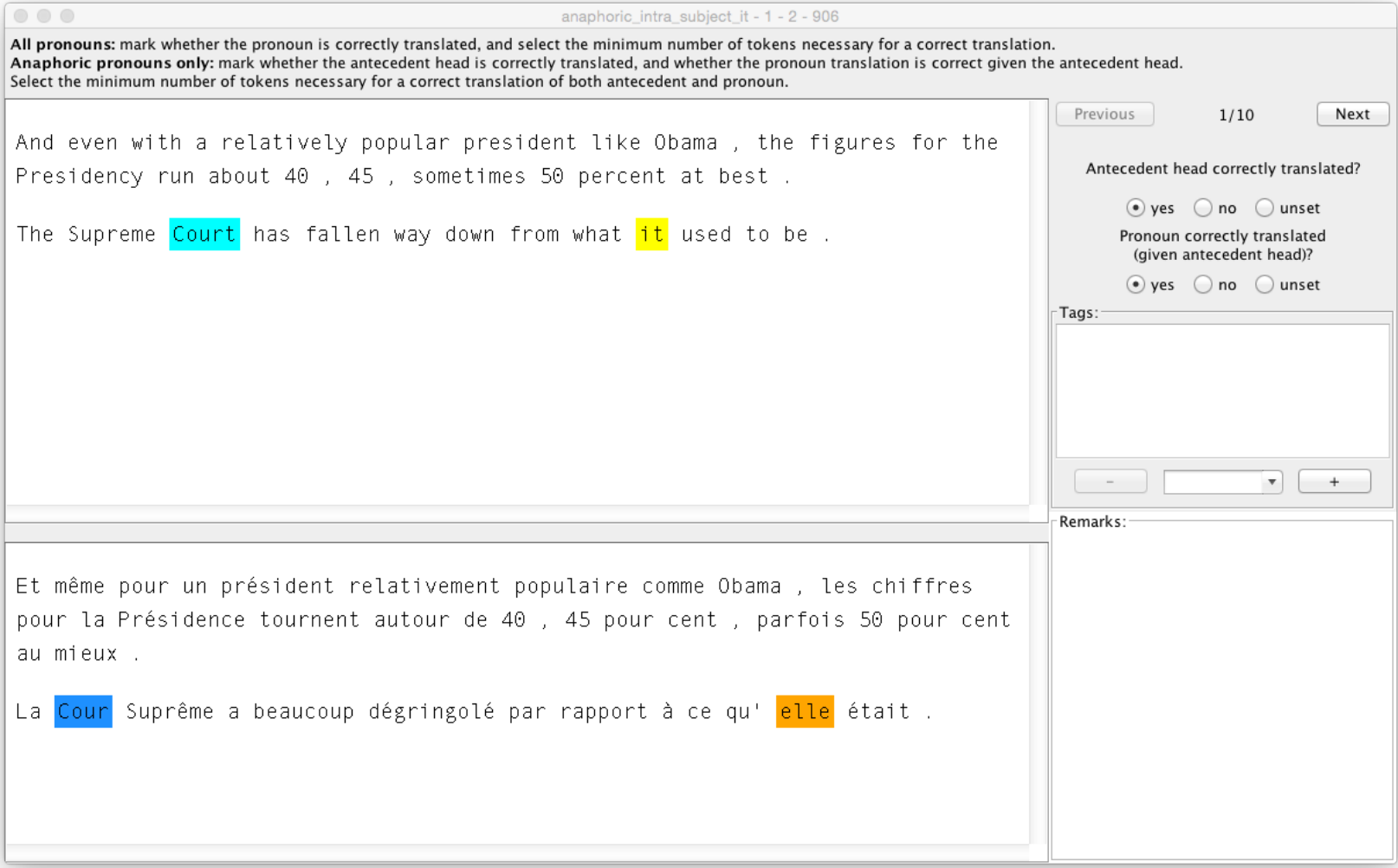
il | elle | ils | elles | ce | cela | il/ce

Other | Bad translation | Discussion required

☐ il ☐ elle ☐ ils ☐ elles ☐ ce ☐ cela

Multiple options possible

0/54 examples annotated.


anaphoric_intra_subject_it - 1 - 2 - 906
All pronouns: mark whether the pronoun is correctly translated, and select the minimum number of tokens necessary for a correct translation.
Anaphoric pronouns only: mark whether the antecedent head is correctly translated, and whether the pronoun translation is correct given the antecedent head.
Select the minimum number of tokens necessary for a correct translation of both antecedent and pronoun.
And even with a relatively popular president like Obama , the figures for the Presidency run about 40 , 45 , sometimes 50 percent at best .
The Supreme Court has fallen way down from what it used to be .
Et même pour un président relativement populaire comme Obama , les chiffres pour la Présidence tournent autour de 40 , 45 pour cent , parfois 50 pour cent au mieux .
La Cour Suprême a beaucoup dégringolé par rapport à ce qu' elle était .
Previous
1/10
Next
Antecedent head correctly translated?
yes
no
unset
Pronoun correctly translated (given antecedent head)?
yes
no
unset
Tags:
-
+
Remarks:

Problems of manual evaluation

- Expensive and time-consuming.
- Requires good language proficiency.
- Gap-filling evaluation causes annotators to miss valid cases.

Task-specific automatic evaluation

- Two automatic reference-based metrics have been proposed:
  - AutoPRF (Hardmeier and Federico, 2010)
  - APT (Miculicich Werlen and Popescu-Belis, 2017)
- Both of them use word alignments and a reference translation.
- AutoPRF is a precision/recall metric based on a BLEU-like clipped count.

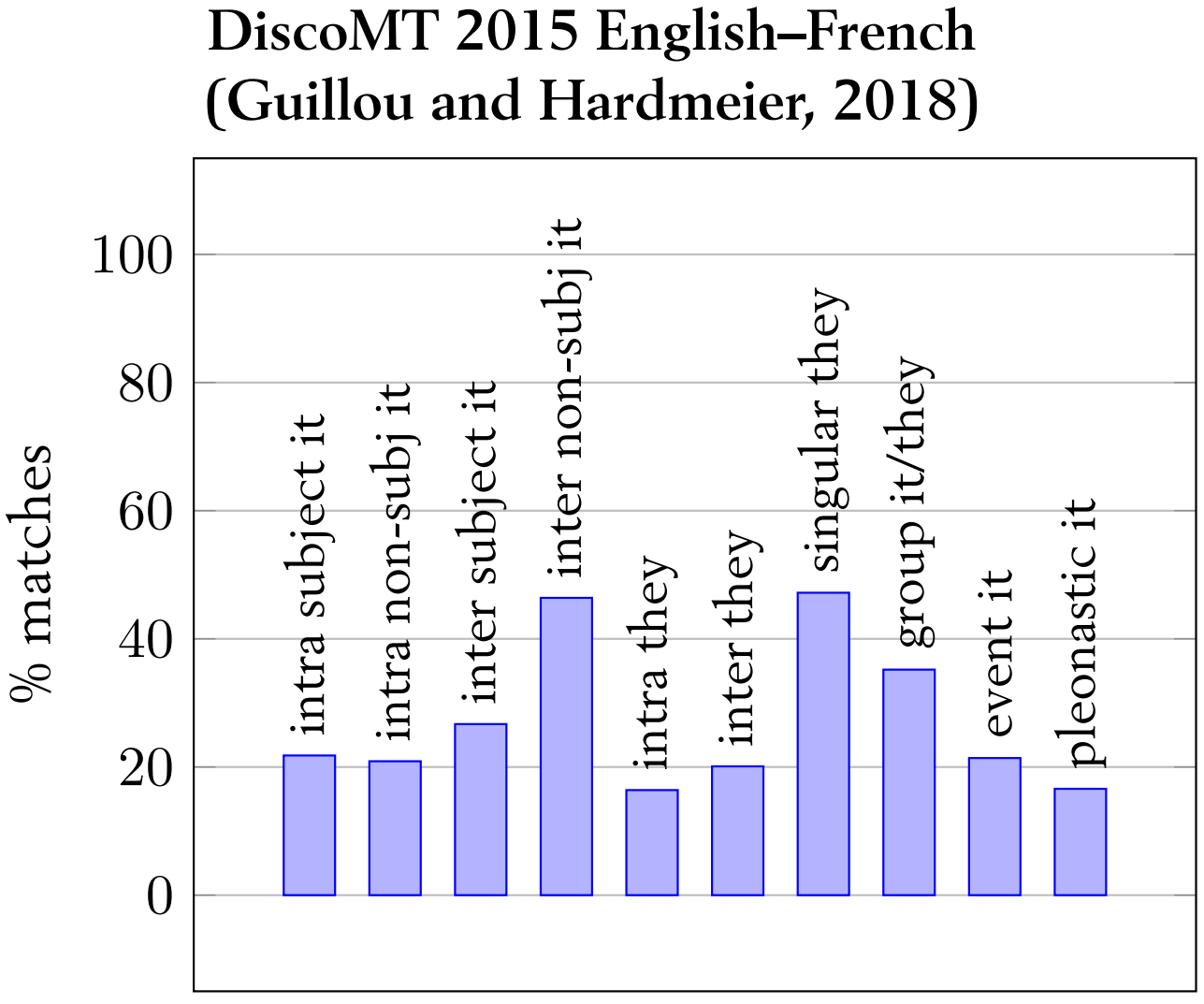
APT: Accuracy of Pronoun Translation

- Create a word alignment
  between the input and the reference translation,
  and the input and the candidate translation.
- Alignments are "improved" using a set of heuristics
  to handle unaligned pronouns.
- Each pronoun is assigned to one of 6 categories:
  1. Identical pronouns
  2. "Equivalent" pronouns according to predefined list
     (e. g., *ce* $\approx$ *il* in French)
  3. Incompatible pronouns
  4. Missing translation in MT output
  5. Missing translation in reference
  6. Missing translation in both
- Scoring with weights for each category.

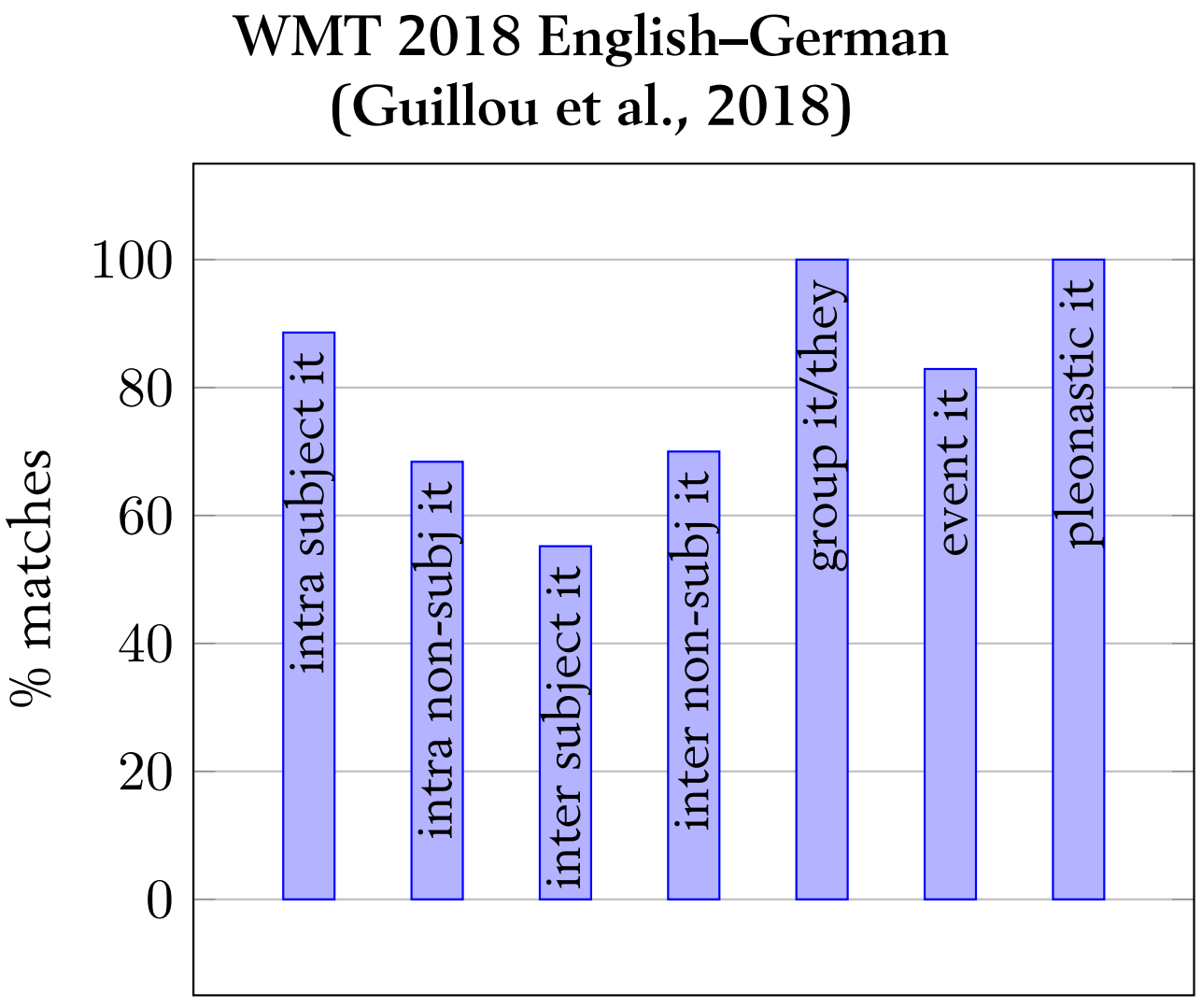
Criticism of automatic pronoun metrics

- Guillou and Hardmeier (2018):
  Detailed comparison with manual evaluation.
- We recommend APT over AutoPRF.
- Alignment heuristics have little effect.
- Pronoun equivalence lists are too simplistic.
  They only work for certain types of pronouns (if at all)
  and introduce errors for other categories – don't use!
- APT works well for the simpler categories,
  but underperforms on some of the categories
  we are most interested in.

Disagreement between human assessment and APT



Validity of APT-like automatic match



Test suites

- Targeted evaluation with limited effort can be achieved with **test suites**, collections of examples covering specific phenomena.
- Recently proposed MT test suites:
  - PROTEST – Pronoun evaluation test suite categorised by function (en-fr, en-de)
  - Isabelle et al. (2017) – General test suite covering various phenomena (en-fr)
  - Bawden et al. (2018) – Anaphoric pronouns and lexical cohesion (en-fr)

Bawden et al. (2018): Anaphoric pronouns and lexical cohesion

- English–French.
- Example blocks consisting of groups of sentence pairs.
- Constructed examples.
- Language modelling task:
  Translations are given, model must choose.
- 50 examples involving coreference and gender agreement.
- 100 examples involving lexical choice
  and consistency/disambiguation.
- Fully automatic evaluation.

Source:

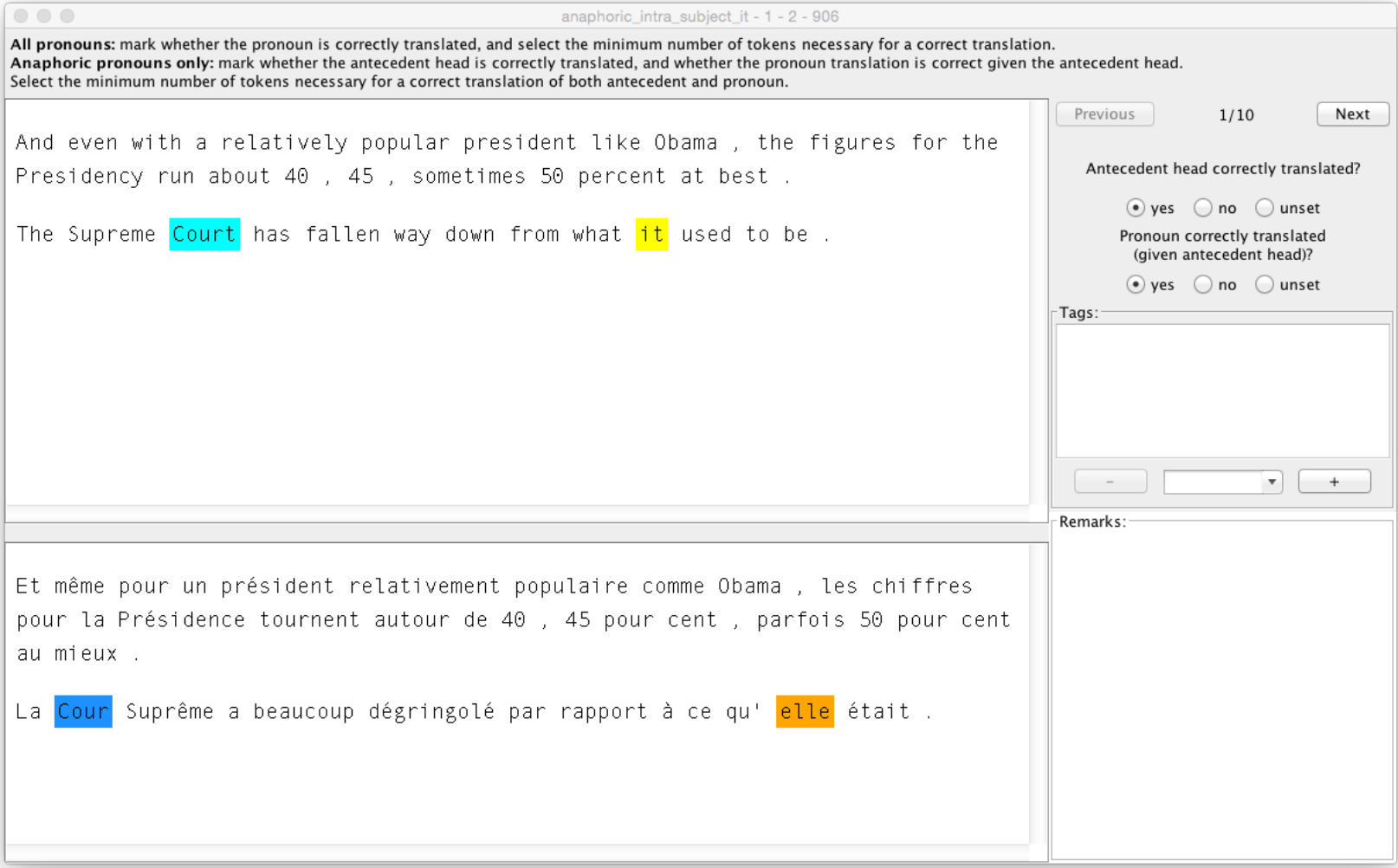
context:	Oh, I hate **flies**. Look, there's another one!
current sent.:	Don’t worry, I'll kill **it** for you.

Target:

1	context:	Ô je déteste les **mouches**. Regarde, il y en a une autre !
	correct:	T'inquiète, je **la** tuerai pour toi.
	incorrect:	T'inquiète, je **le** tuerai pour toi.
2	context:	Ô je déteste les **moucherons**. Regarde, il y en a un autre !
	correct:	T'inquiète, je **le** tuerai pour toi.
	incorrect:	T'inquiète, je **la** tuerai pour toi.
3	context:	Ô je déteste les **araignées**. Regarde, il y en a une autre !
	semi-correct:	T'inquiète, je **la** tuerai pour toi.
	incorrect:	T'inquiète, je **le** tuerai pour toi.
4	context:	Ô je déteste les **papillons**. Regarde, il y en a un autre !
	semi-correct:	T'inquiète, je **le** tuerai pour toi.
	incorrect:	T'inquiète, je **la** tuerai pour toi.

PROTEST Test Suite

- English–French (Guillou and Hardmeier, 2016) and English–German (Guillou et al., 2018).
- Examples selected from corpus data (TED talks).
- Machine translation task: Only source is given.
- 200 tokens of *it* and *they*, stratified by pronoun category.
- EN-FR version also includes 50 tokens of *you*.
- Semi-automatic or manual evaluation with graphical user interface (Hardmeier and Guillou, 2016).


anaphoric_intra_subject_it - 1 - 2 - 906
All pronouns: mark whether the pronoun is correctly translated, and select the minimum number of tokens necessary for a correct translation.
Anaphoric pronouns only: mark whether the antecedent head is correctly translated, and whether the pronoun translation is correct given the antecedent head.
Select the minimum number of tokens necessary for a correct translation of both antecedent and pronoun.
And even with a relatively popular president like Obama , the figures for the Presidency run about 40 , 45 , sometimes 50 percent at best .
The Supreme Court has fallen way down from what it used to be .
Et même pour un président relativement populaire comme Obama , les chiffres pour la Présidence tournent autour de 40 , 45 pour cent , parfois 50 pour cent au mieux .
La Cour Suprême a beaucoup dégringolé par rapport à ce qu' elle était .
Previous
1/10
Next
Antecedent head correctly translated?
yes
no
unset
Pronoun correctly translated (given antecedent head)?
yes
no
unset
Tags:
-
+
Remarks:

On test suites and cherry picking

A common method of presenting work on discourse in MT:

- Run experiments.
- Find improvements in some automatic score you like.
- In paper, present automatic scores and "qualitative analysis", showing 3 examples from your output.
- Draw long-reaching conclusions about all the cool things your system has learnt about discourse.
- Publish and enjoy.

On test suites and cherry picking

A common method of presenting work on discourse in MT:

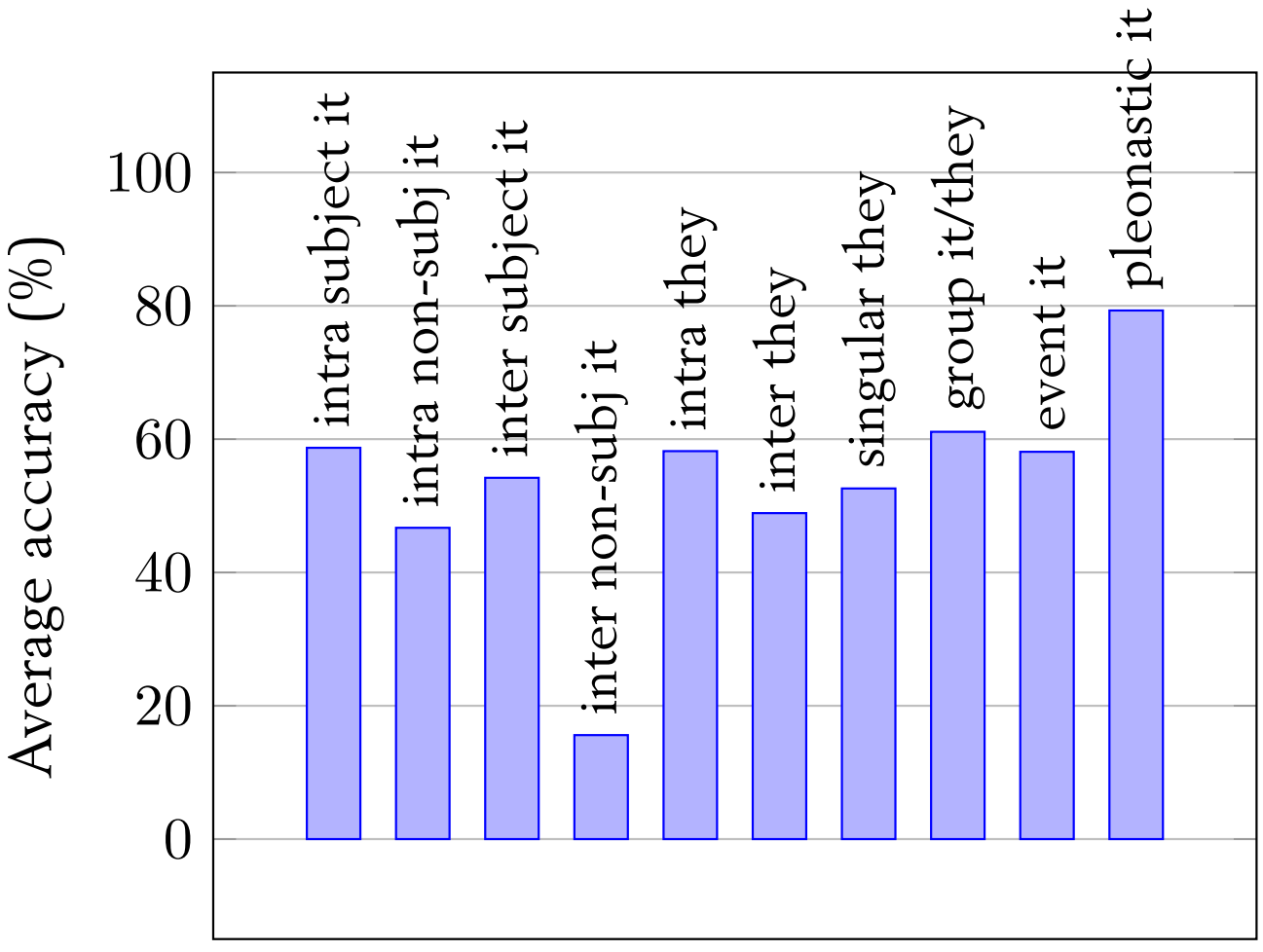
- Run experiments.
- Find improvements in some automatic score you like.
- In paper, present automatic scores and "qualitative analysis", showing 3 examples from your output.
- Draw long-reaching conclusions about all the cool things your system has learnt about discourse.
- Publish and enjoy.
- Cherry-picked examples aren't worth very much.
- Your argument becomes a lot more credible if you select your examples beforehand (ad hoc test suite).

So how good is MT at the moment?

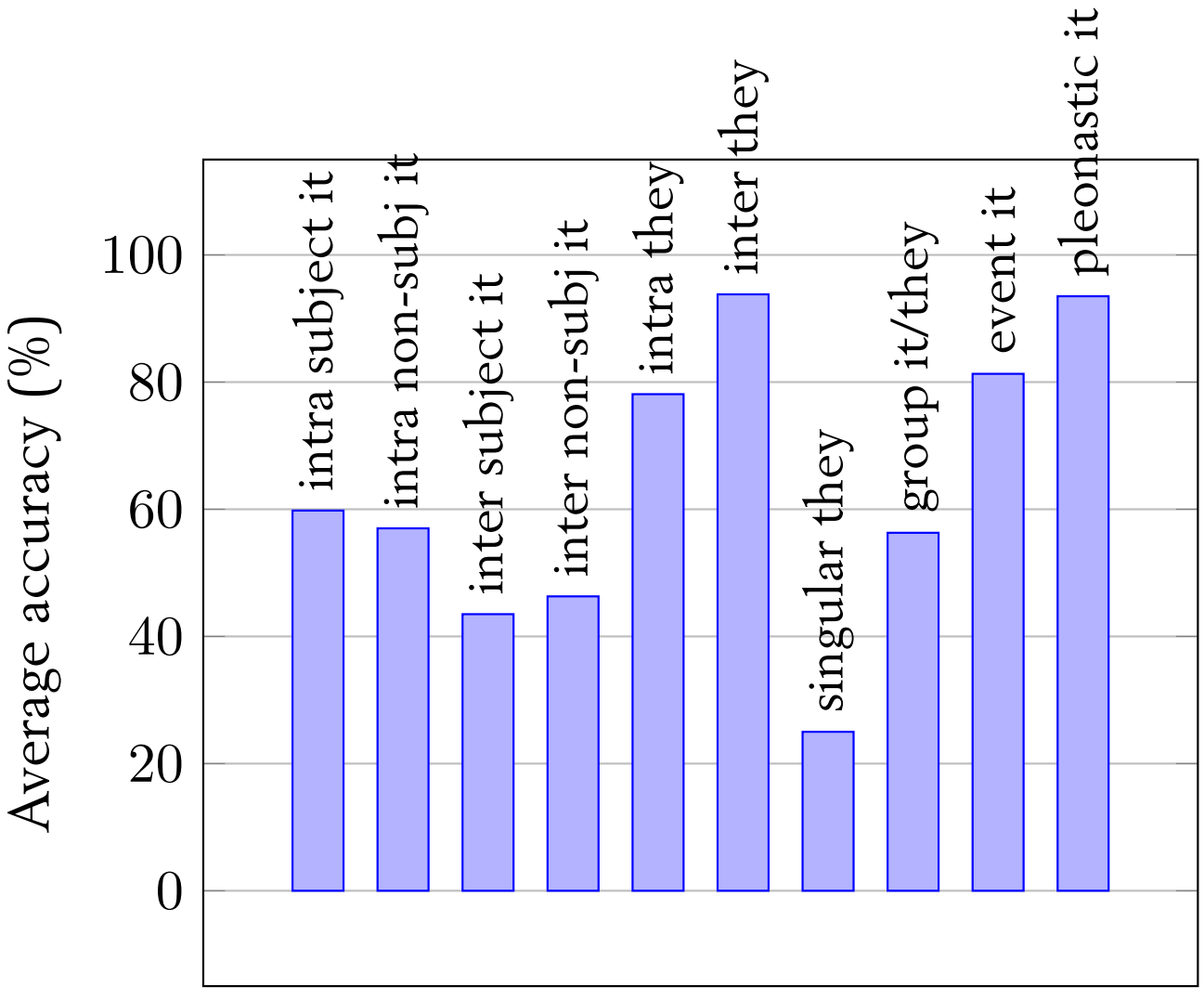
Two evaluations with the PROTEST test suite:

- DiscoMT 2015 EN-FR (Hardmeier and Guillou, 2018):
  - 5 SMT systems
  - 1 rule-based MT system
  - 2 recurrent NMT systems
  - 1 Transformer system with context encoder
- WMT 2018 EN-DE (Guillou et al., 2018):
  - 10 shared task submissions (all neural, mostly Transformer)
  - 6 anonymous online systems

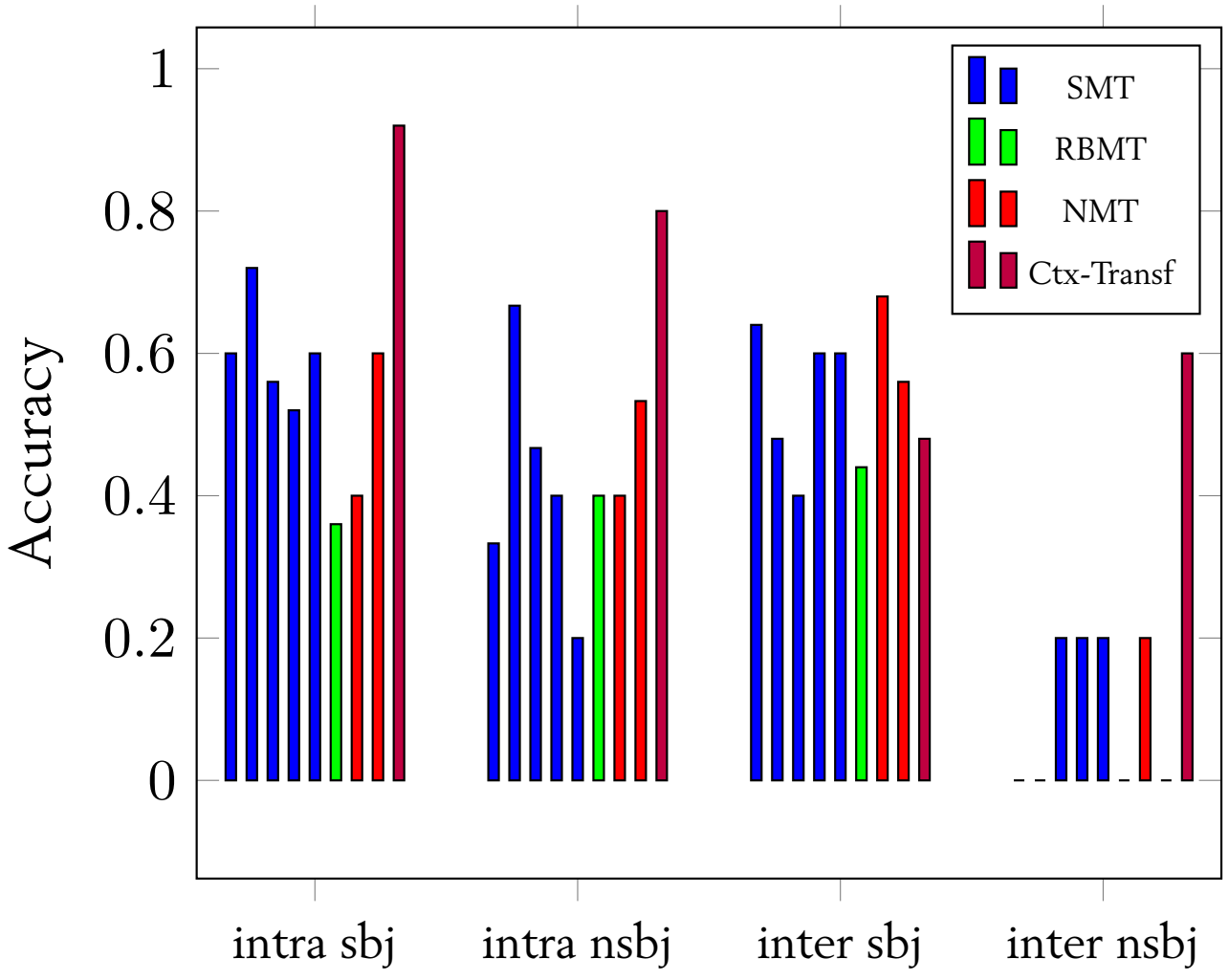
DiscoMT 2015 EN-FR



WMT 2018 EN-DE



DiscoMT 2015 EN-FR



Recent approaches to cross-sentence NMT

- Flurry of publications in 2017 and 2018,
  check EMNLP and WMT.
- General approach: Adding one or a few sentences of
  cross-sentence context to a standard NMT system,
  hope for the best.
- Everything comes in an RNN and a self-attention flavour.
- No common benchmarks,
  very difficult to compare performance.

Types of context

Context can be provided

- on the source side:
  Providing additional information
  to disambiguate the input.
- on the target side:
  Keep track of translations in other sentences
  to improve consistency
  (but cave Carpuat and Simard, 2012).
- on both sides.

Simple preprocessing

Tiedemann and Scherrer (2017), DE-EN

- Context is fed into an NMT system as additional tokens.
- Simple manipulations of the input data,
  no changes to system implementation.
- Source and target side context, different setups.

Encoder for previous sentence

- Jean et al. (2017), EN-FR/EN-DE:
  - Additional RNN encoder for previous source sentence.
  - Winning system of the DiscoMT 2017 pronoun prediction task.
- Voita et al. (2018), EN-RU:
  - Transformer model with an additional encoder for the previous sentence.
  - Evidence of specific improvements for coreference translation (pronoun-specific BLEU evaluation and analysis of attention weights).
  - PROTEST evaluation on EN-FR confirm improvements for intra-sentential anaphora,
    but cross-sentence cases remain problematic (Hardmeier and Guillou, 2018).

Hierarchical context summarisation

- Wang et al. (2017), EN-ZH:
  - Hierarchical RNN summarising a fixed number of context sentences.
  - Used for decoder initialisation and as additional context.
- Miculicich et al. (2018), ZH-EN/ES-EN:
  - Hierarchical attention networks summarising a fixed number of context sentences.
  - Demonstrates improvements using automatic APT metric.

Memory networks

- Tu et al. (2018), ZH-EN:
  - Neural cache data structure that is indexed by a source context vector and stores target-side representations.
- Maruf and Haffari (2018), FR-EN/DE-EN/EE-EN:
  - Memory network to condition the generation of a word not only on the current sentence, but on the whole document.
  - Computationally very demanding, evaluated on very small training sets only.

Relevance

- NMT is becoming better and better at the sentence level, but discourse problems remain.
- MT output that is indistinguishable from human translation when considered sentence by sentence can still be worse when you look at enough context (Läubli et al., 2018).
- Clearly, research on discourse-level MT is becoming topical.
- Results so far are encouraging, but rather preliminary.

Some considerations

- When you create corpora, don't discard document information.
- When you plan MT evaluations, evaluate in context.
- Most work so far has considered
  very limited additional context,
  but is there a way to go truly global?

MT and Artificial Intelligence

- Ultimately, coherence is about mental processes in the reader. . .
- . . . and translation is really about communication.
- Can we tackle translation
  as a form of language understanding
  instead of pattern matching and transduction?