HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

Hes·so
Haute Ecole Spécialisée
de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences and Arts
Western Switzerland

Context in NMT evaluation: methods and implications for MT

Andrei Popescu-Belis

School of Management and Engineering Vaud (HEIG-VD)

University of Applied Sciences of Western Switzerland (HES-SO)

Yverdon-les-Bains

SMART Select Workshop, Luxembourg, 19 November 2019

DOMAT GRANT 175693

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

Does document-level MT improve document-level quality?

Two different goals

1. Design MT methods which are able to consider context
    - Context ≈ features involving long-range dependencies (>> phrase)
        - inter-clause (intra-sentential), inter-sentential, or document-level

2. Evaluate document-level aspects of quality
    - *“Do you mean document-level BLEU ? Well, not only.”*
    - Correct translation of phenomena that are hard to translate without context
        - discourse-level phenomena ≈ those that profit from pragmatic knowledge

Are the two goals correlated?

- Many recent NMT studies attempt to answer questions such as:

$$\text{Does our} \left\{ \begin{array}{l} \text{document-level} \\ \text{local-level} \\ \text{hybrid/modular} \end{array} \right\} \text{model improve translation} \left\{ \begin{array}{l} \text{at the document level ?} \\ \text{at the word/sentence level ?} \\ \text{at an unspecified level ?} \end{array} \right.$$

- This talk: how do we measure document-level quality?
  - Taxonomize and exemplify types of *quality measures*, *not MT models*

Outline

1. A taxonomy of MT evaluation: remember FEMTI?
2. Measures of document-level quality for recent NMT models
   - BLEU as a general indicator of quality
   - Grammatical/lexical quality and contrastive pairs (a parenthesis)
   - Measuring semantic and discourse phenomena
     - lexical choice: WSD and non-WSD
     - pronouns and coreference
     - discourse structure and connectives

- Reference list available at https://arxiv.org/abs/1901.09115

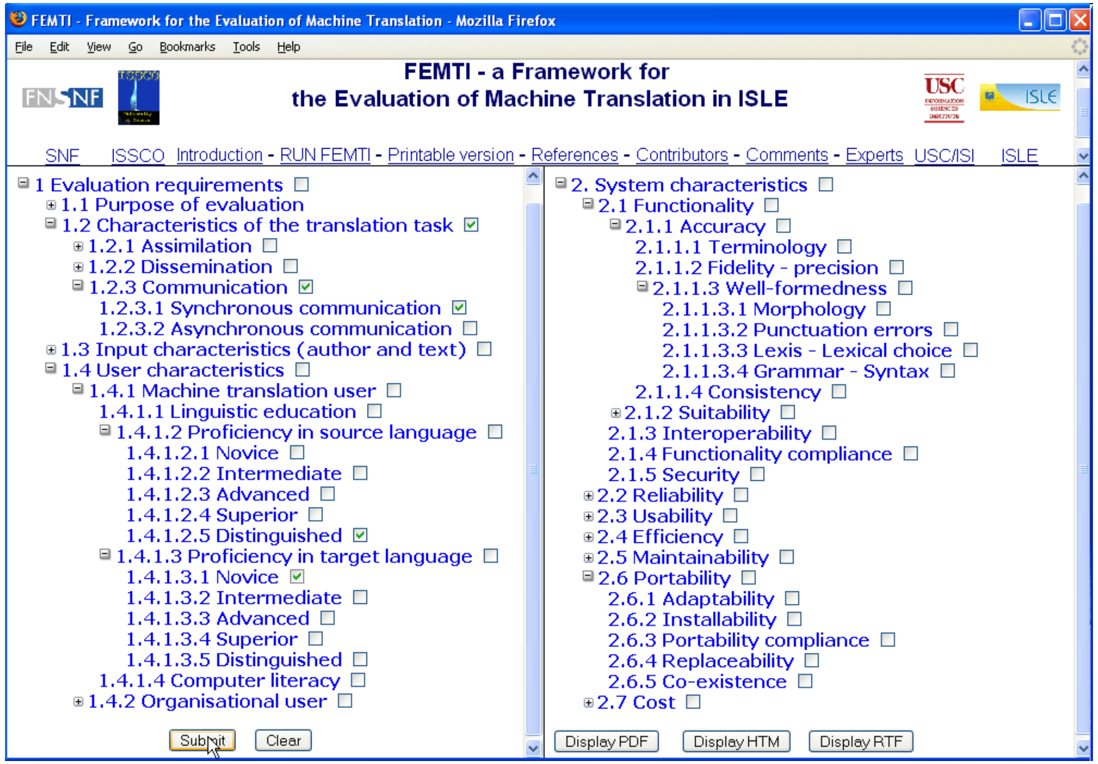
FEMTI

Hovy, King, and Popescu-Belis (2003)

and several other papers

FEMTI: Framework for the Evaluation of Machine Translation in ISLE (EU project in the early 2000s)

- ISLE: International Standards for Language Engineering
  - collected knowledge from the MT community, 100+ evaluation metrics over the past 30 years
  - inspired by ISO/IEC standards
    1. classification of contexts of use
    2. classif. of quality characteristics
    3. mechanism to generate context-based evaluation plans


FEMTI - Framework for the Evaluation of Machine Translation - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
FN SNF
ISSCO
USC
ISLE
SNF ISSCO Introduction - RUN FEMTI - Printable version - References - Contributors - Comments - Experts USC/ISI ISLE
Submit
Clear
Display PDF
Display HTM
Display RTF

FEMTI - a Framework for the Evaluation of Machine Translation in ISLE

- ⊟ 1 Evaluation requirements □
  - ⊞ 1.1 Purpose of evaluation
  - ⊟ 1.2 Characteristics of the translation task ☑
    - ⊞ 1.2.1 Assimilation □
    - ⊞ 1.2.2 Dissemination □
    - ⊟ 1.2.3 Communication ☑
      - 1.2.3.1 Synchronous communication ☑
      - 1.2.3.2 Asynchronous communication □
  - ⊞ 1.3 Input characteristics (author and text) □
  - ⊟ 1.4 User characteristics □
    - ⊟ 1.4.1 Machine translation user □
      - 1.4.1.1 Linguistic education □
      - ⊟ 1.4.1.2 Proficiency in source language □
        - 1.4.1.2.1 Novice □
        - 1.4.1.2.2 Intermediate □
        - 1.4.1.2.3 Advanced □
        - 1.4.1.2.4 Superior □
        - 1.4.1.2.5 Distinguished ☑
      - ⊟ 1.4.1.3 Proficiency in target language □
        - 1.4.1.3.1 Novice ☑
        - 1.4.1.3.2 Intermediate □
        - 1.4.1.3.3 Advanced □
        - 1.4.1.3.4 Superior □
        - 1.4.1.3.5 Distinguished □
      - 1.4.1.4 Computer literacy □
    - ⊞ 1.4.2 Organisational user □

- ⊟ 2. System characteristics □
  - ⊟ 2.1 Functionality □
    - ⊟ 2.1.1 Accuracy □
      - 2.1.1.1 Terminology □
      - 2.1.1.2 Fidelity - precision □
      - ⊟ 2.1.1.3 Well-formedness □
        - 2.1.1.3.1 Morphology □
        - 2.1.1.3.2 Punctuation errors □
        - 2.1.1.3.3 Lexis - Lexical choice □
        - 2.1.1.3.4 Grammar - Syntax □
      - 2.1.1.4 Consistency □
    - ⊞ 2.1.2 Suitability □
    - 2.1.3 Interoperability □
    - 2.1.4 Functionality compliance □
    - 2.1.5 Security □
  - ⊞ 2.2 Reliability □
  - ⊞ 2.3 Usability □
  - ⊞ 2.4 Efficiency □
  - ⊞ 2.5 Maintainability □
  - ⊟ 2.6 Portability □
    - 2.6.1 Adaptability □
    - 2.6.2 Installability □
    - 2.6.3 Portability compliance □
    - 2.6.4 Replaceability □
    - 2.6.5 Co-existence □
  - ⊞ 2.7 Cost □

https://www.isi.edu/natural-language/mteval/

Excerpt from FEMTI

- **2.2.1 Functionality**
  - **2.2.1.1 Suitability**
    - **2.2.1.1.1 Target-language only**
      - **2.2.1.1.1.1 Readability (or: fluency, intelligibility, clarity)**
      - **2.2.1.1.1.2 Comprehensibility**
      - **2.2.1.1.1.3 Coherence**
      - **2.2.1.1.1.4 Cohesion**
    - **2.2.1.1.2 Cross-language / contrastive**
      - **2.2.1.1.2.1 Coverage of corpus-specific phenomena**
      - **2.2.1.1.2.2 Style**
  - **2.2.1.2 Accuracy**
    - **2.2.1.2.1 Fidelity**
    - **2.2.1.2.2 Consistency**
    - **2.2.1.2.3 Terminology**
  - **2.2.1.3 Wellformedness**
    - **2.2.1.3.1 Punctuation**
    - **2.2.1.3.2 Lexis / lexical choice**
    - **2.2.1.3.3 Grammar / syntax**
    - **2.2.1.3.4 Morphology**

- Coherence: “the degree to which the reader can describe the role of each individual sentence (or group of sentences) with respect to the text as a whole. Theories such as Rhetorical Structure Theory attempt to formalize coherence.”
  - Metric: e.g. by counting the total number of sentences in MT output to which RST labels can be assigned
- Cohesion: “refers to lexical chains and other elements – for example lexical chains, anaphora, ellipsis – that link individual units across sentences.
  - Metric: does the system render cohesive units appropriately for the target language?
- Style: “subjective evaluation of the correctness of the style (or register) of each sentence”

Types of metrics and datasets

- Reference-based (“objective”)
  - automatically measure word-based similarity with a reference translation (all words or subset)
- Human-based (“subjective”)
  - human assessment of correctness, using the source and possibly a reference translation
- Test suite/challenge set
  - a test set focused on specific phenomena, with reference-based or human-based metrics
- Contrastive pairs
  - given two translation options, the system is asked to rank them by their likelihood
- Evaluation in use: use MT output for other tasks (IR, QA, etc.)

Document-level quality measures

Global document-level quality: BLEU or humans?

- Impact on NMT of degraded context (Kim et al. 2019)
  - contextual NMT (Transformer): concatenate sentences on source/target side
  - experiments: remove stopwords, most frequent words, keep only some POS
  - consider variation of BLEU scores
  - ➢ context is mostly useful to provide a general representation of the topic

- Agrawal et al. (2018): BLEU improves by concatenating src/tgt sentences (fewer data and larger improvement)

- Human judges rating overall quality at the text level rather than the sentence level (Läubli et al. 2018, Toral et al. 2018)
  - reassessment of Hassan et al.'s (2018) claim of human parity with Bing NMT
  - ➢ when texts are rated by professional translators, the difference between humans and NMT becomes significant

- Low power of statistical tests for document-level human ratings (Graham et al. 2019)

Grammatical/lexical quality (a parenthesis)

- Initial stages of NMT (2016-2017)
  - analyses of NMT output based on taxonomies of grammatical and lexical errors
  - error counts obtained from human judges

- Bentivogli et al. (2016)
  - lexical errors (wrong lemma); morphological errors (correct lemma but wrong form); word order errors

  ➢ NMT: 20% fewer lex./morph. mistakes than SMT but sometimes skipped negations 💣

- Other studies
  - Castilho et al. (2017) add fluency, adequacy
  - Popovic (2017)
  - Toral and Sánchez-Cartagena (2017)
  - Klubička et al. (2018) with MQM taxonomy

➢ Typically rating hundreds of sentences

➢ Superiority of NMT at the local level

➢ Not explicitly at the document level

  - some lexical errors could be "contextual"

Assessing grammatical/lexical quality with contrastive pairs or test suites

- LingEval97 (Sennrich 2017)
  - 97,000 EN/DE sentence pairs
    - correct translation + wrong counterpart obtained by rule-based changes of REF (gender of determiners, verb number or particle, polarity, etc.)
      - SRC: *Prague Stock Market falls to minus by the end of the trading day.*
      - REF: *Die Prager Börse stürzt gegen Geschäftsschluss ins Minus.*
      - POLARITY: *ins >> nicht ins*
      - NUMBER: *stürzt >> stürzen*
  - systems rank translations by likelihood

- Challenge Set (Isabelle et al. 2017)
  - 108 sentences with EN/FR divergencies
    - morphosyntactic (e.g. agreement), syntactic (e.g. position of clitic pronouns), lexico-syntactic (e.g. double objects)
      - SRC: *Mary manque beaucoup à John.*
      - REF: *John misses Mary a lot.*
  - human judges evaluate translations

Lexical choice: WSD vs. non-WSD errors

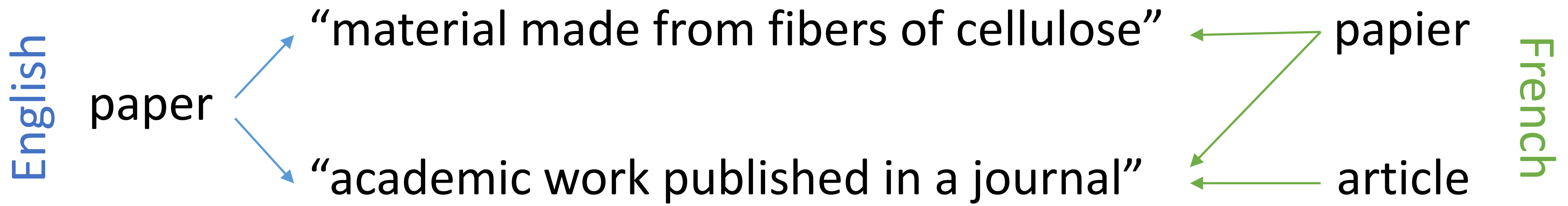
Are lexical errors semantic or discursive?

- SRC: The method finds a minimum spanning tree if the graph is connected. But if the graph is not connected, then it finds a minimum spanning forest.
- NMT: La méthode recherche un spanning tree minimum si le graphique est connecté. Mais si le graphique n'est pas relié, il trouve une forêt minimale couvrant.
- REF: La méthode trouve un arbre couvrant minimal si le graphe est connecté. Mais si le graphe n'est pas connecté, elle trouve une forêt couvrante minimale.

➢ Categorizing an MT error as semantic or discursive often involves a hypothesis on the cause of an error, or on the features that would enable a system to avoid it.

❖ Discursive ≈ which cannot be translated accurately above chance without considering previous clause(s) or sentence(s)

Mistranslation of word senses (1)



English

paper

“material made from fibers of cellulose”

“academic work published in a journal”

papier

article

French

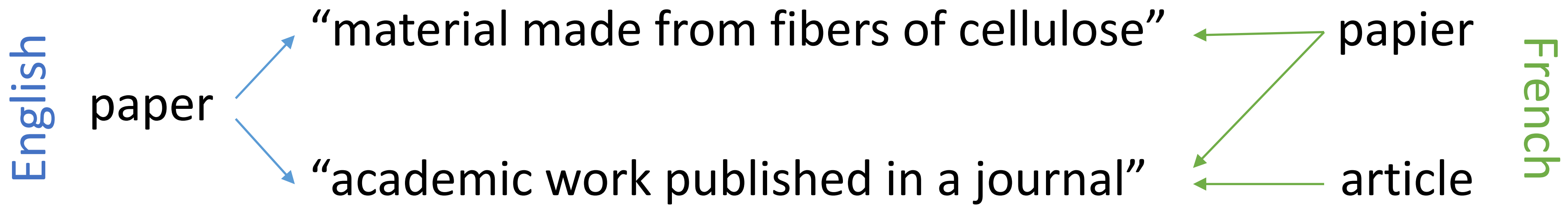
Paper is a thin material produced from cellulose pulp. Papers are essential in legal documentation.

Le papier est un matériau fin fabriqué à partir de pâte de cellulose. Les articles sont essentiels dans la documentation juridique.

Incoherent translation: the meaning of papers (2nd occ.) is misunderstood, a word sense disambiguation error

Maybe the system should have looked at the surrounding words?

Mistranslation of word senses (2)



English

paper

“material made from fibers of cellulose”

“academic work published in a journal”

papier

article

French

There are ten different types of scientific papers. […] Papers that carry specific objectives are: …

Il existe dix types d'articles scientifiques différents. […] Les papiers qui ont des objectifs spécifiques sont : …

Inconsistent translation, the 2nd occurrence of papers should have been rendered by the same word (but this is not a WSD error)

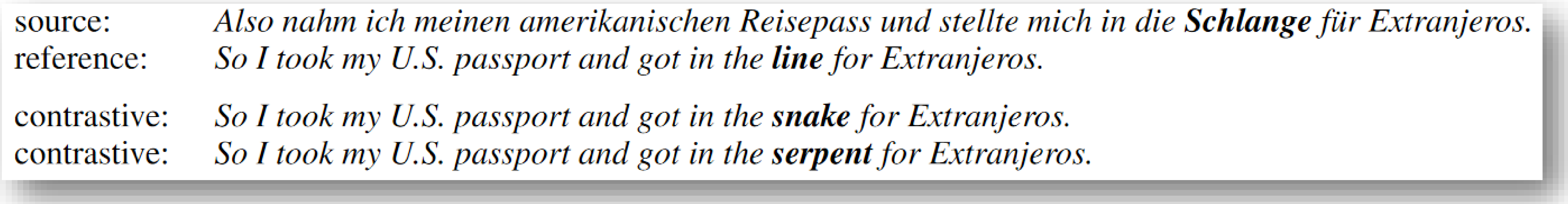
Maybe the system could have looked at the first occurrence?

Evaluating lexical errors with contrastive pairs (1)

- ContraWSD (Rios et al. 2017)
  - same principle as Lingeval97
  - DE/EN and FR/EN
  - contrastive pairs to evaluate translation of polysemous words
  - for 80 word senses, generate several wrong translations by replacing the target word with other observed translations of the word: avg. 90 sentences/sense
  - system must rank alternatives

➢results on DE/EN

- Nematus (Sennrich et al. 2017): 70%
- sense-aware system: 70%



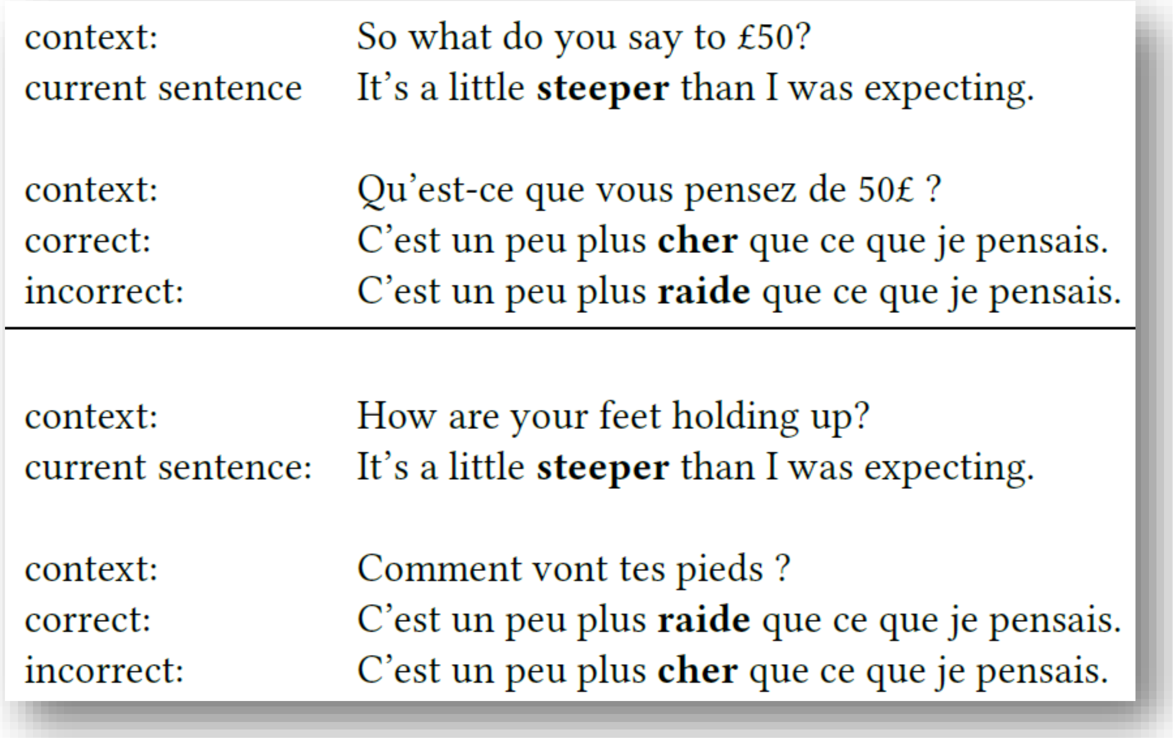
source:	*Also nahm ich meinen amerikanischen Reisepass und stellte mich in die **Schlange** für Extranjeros.*
reference:	*So I took my U.S. passport and got in the **line** for Extranjeros.*
contrastive:	*So I took my U.S. passport and got in the **snake** for Extranjeros.*
contrastive:	*So I took my U.S. passport and got in the **serpent** for Extranjeros.*

Contrastive pair of translations (Rios et al. 2017, Table 1, p. 14)

Evaluating lexical errors with contrastive pairs (2)

- Lexical choice set (Bawden 2018)
  - 100 couples of pairs, so that non-contextual MT can only get one correct and one wrong
  - WSD errors ≈ coherence errors (they lead to “incoherent” output)
  - non-WSD errors ≈ cohesion errors

  ➢ best results
    - multi-encoder ‘S-HIER-TO-2’ 57%
    - concatenation ‘2-TO-1’ 53%



context:	So what do you say to £50?
current sentence	It’s a little **steeper** than I was expecting.
context:	Qu’est-ce que vous pensez de 50£ ?
correct:	C’est un peu plus **cher** que ce que je pensais.
incorrect:	C’est un peu plus **raide** que ce que je pensais.
context:	How are your feet holding up?
current sentence:	It’s a little **steeper** than I was expecting.
context:	Comment vont tes pieds ?
correct:	C’est un peu plus **raide** que ce que je pensais.
incorrect:	C’est un peu plus **cher** que ce que je pensais.

Couple of pairs for testing WSD (Bawden 2018, p. 159)

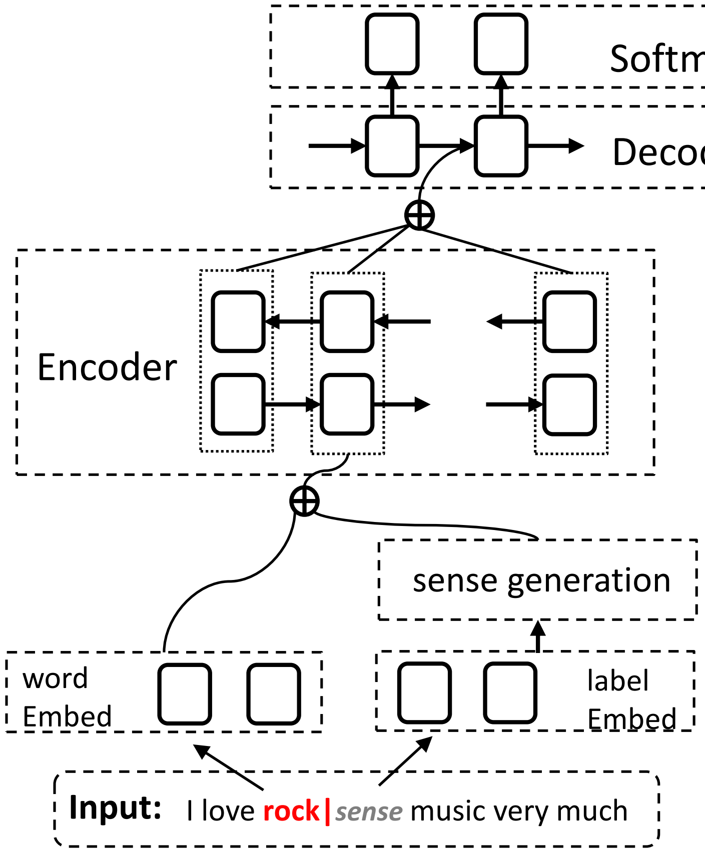
Limitations of contrastive pairs

- They require access to the probability estimates of pairs of source and target sentences from the evaluated system (for ranking alternatives)
  - Easy to obtain from one’s own NMT system, but impossible from online NMT

- They do not guarantee that if a system is better than another for ranking pairs of candidate target sentences, it will also be better when it comes to *finding* the correct target when only the source is given

- Not quite naturally-occurring texts

Example: sense-aware NMT (Pu et al. 2018)



E = [E_{rock} ; E_{sense}]

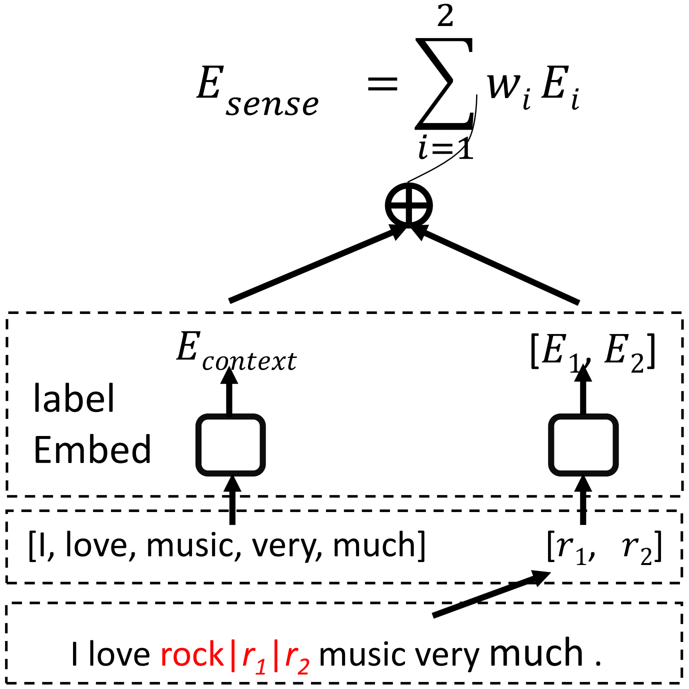
- Four sense embedding models

  TOP: use sense found by WSD

  AVG: weighted average of senses

  ATT: attention-based sense weights, computed dynamically during encoding

  ATT$_{\text{ini}}$: ATT model with source word vectors initialized using word2vec



$$E_{sense} = \sum_{i=1}^{2} w_i E_i$$

$E_{context}$
$[E_1, E_2]$
label
Embed
[I, love, music, very, much]
$[r_1, r_2]$
I love rock|r_1|r_2 music very much .

Evaluation measures on EN/FR (Pu et al. 2018)

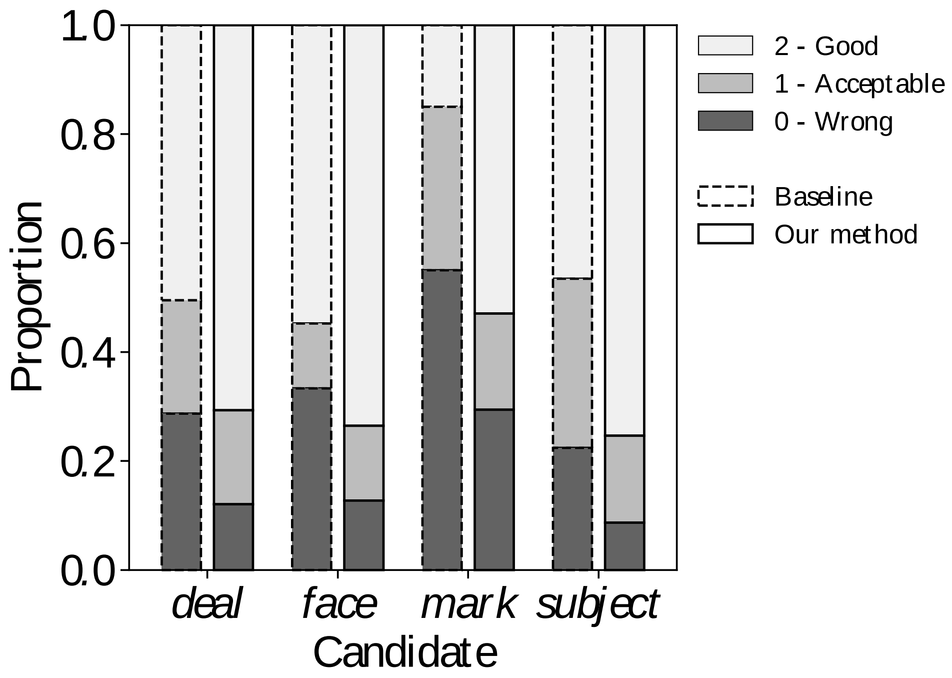
- “Objective” evaluation with BLEU

Model	BLEU (w. Δ)
Baseline	34.6
TOP	34.5 (-0.1)
AVG	35.2 (+0.6)
ATT	35.3 (+0.7)
ATT$_{ini}$	**35.8 (+1.2)**

		Baseline	
		Correct	Incorrect
ATT$_{ini}$	Correct	134,552	17,145
	Incorrect	10,551	101,228

- Identity with reference restricted to nouns and verbs with sense labels

- “Subjective” evaluation: 4 words
  - humans compare baseline with ATT$_{ini}$



Evaluation of pronoun translation

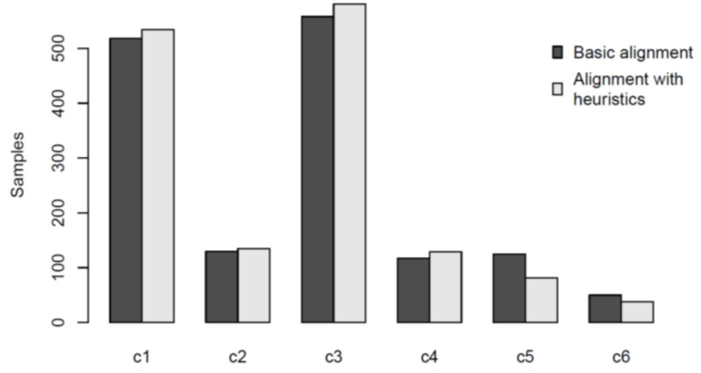
APT Metric: Accuracy of Pronoun Translation

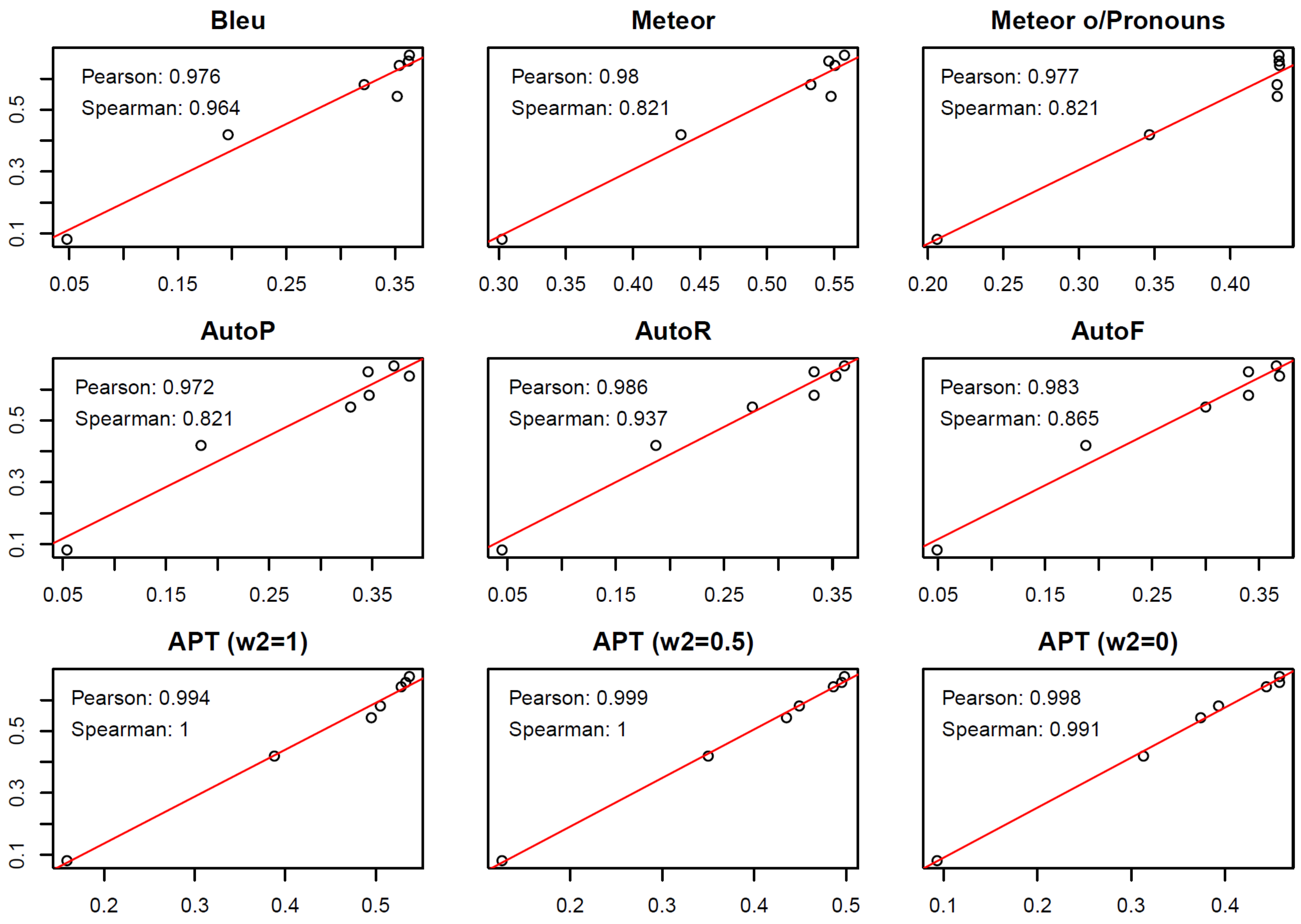
(Miculicich Werlen & Popescu-Belis 2017)

- Compare NMT translations of pronouns to human reference (on EN/FR)
  - requires word alignment (GIZA & heuristics)
  - accepts some variation of pronoun choice (e.g. *it is difficult* → *il / c' est difficile*)

- Limitation of reference-based metrics (Guillou & Hardmeier 2018)
  - ➢ different translations can be equally acceptable for a pronoun, depending on the lexical choice for its antecedent

Cases	Score	Meaning
1: Equal	1	Correct translation
2: Equivalent	tunable	Correct, partially correct or incorrect translation
3: Different	0	Incorrect translation
4: Not translated in candidate	0	Incorrect translation
5: Not translated in reference	0	Incorrect translation
6: Not translated in cand. & ref.	tunable	Correct, partially correct or incorrect translation





Correlation between manual evaluation (vertical) and various reference-based metrics (horizontal) for pronoun evaluation (from Miculicich Werlen and Popescu-Belis (2017))

PROTEST and its used in shared tasks (1)

- PROTEST (Guillou and Hardmeier, 2016)
  - test suite with 250 pronouns and their reference translations
    - based on ParCor annotation guidelines for pronoun status and antecedent (Guillou et al., 2014)
  - identity between a candidate and reference pronoun translation is scored automatically, but each difference is submitted to a human judge

- Shared tasks

  1. Pronoun translation
  2. Pronoun prediction, given the source and a lemmatized reference with deleted pronouns

  - both were tried at DiscoMT 2015 (Hardmeier et al., 2015), but only the second one was continued at WMT 2016 and DiscoMT 2017 (Guillou et al., 2016; Loáiciga et al., 2017)

PROTEST and its used in shared tasks (2)

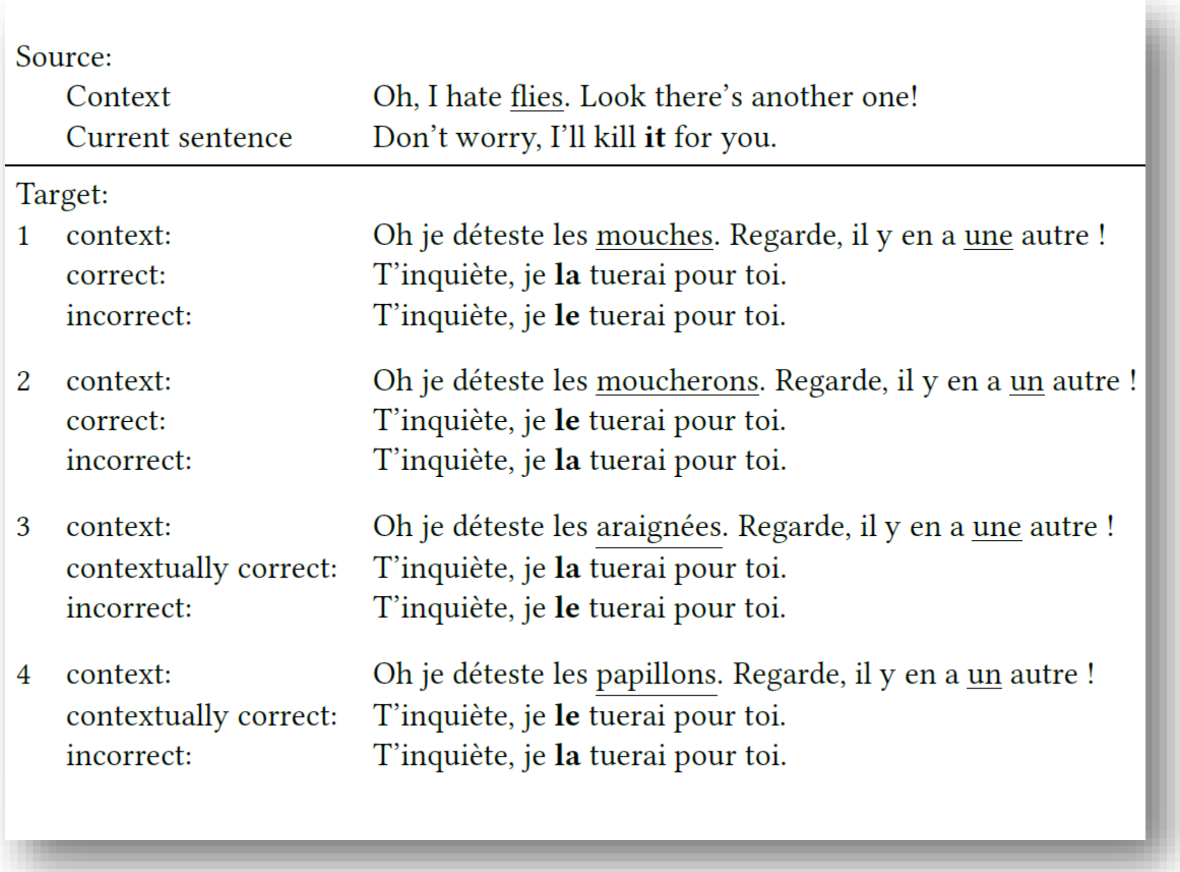
- Shared task at WMT 2018 (Guillou et al., 2018)
  - 16 systems from the EN/DE news task; PROTEST style, 200 occurrences of *it* and *they*
  - ➢ 50% of the systems translate correctly more than 145 pronouns (the best one, Marian, reaches 157)
    - scores correlate with BLEU (r = 0.91) and APT (r = 0.89)

- Same method on EN/FR (Hardmeier & Guillou, 2018) with 250 occurrences of *it* and *they*
  - ➢ average score over 9 systems: 160/250 | best score for the system by Voita et al. (2018): 199/250 (good on non-referential or intra-sentential anaphoric *it* and *they*, but not on inter-sentential ones)

- PROTEST scores from Scherrer et al. (2019) for EN/DE NMT with concatenated sentences
  - ➢ improvement on subtitles (from 91 to 100/200), but no improvement on news (108/200)

Contrastive pairs (1)

- Pronoun set (Bawden et al. 2018)
  - 100 blocks, personal and possessives
  - generate 4 alternatives for the translation of each antecedent: (a) reference; (b) correct but opposite gender; (c, d) inaccurate (F/M)
  - contrastive pair: F/M pronoun
  - expected rankings: for (a) and (b) the correct gender; for (c) and (d) gender of the inaccurate translation (“contextually correct”)



Source:		
	Context	Oh, I hate flies. Look there’s another one!
	Current sentence	Don’t worry, I’ll kill **it** for you.
Target:		
1	context:	Oh 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:	Oh 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:	Oh je déteste les araignées. Regarde, il y en a une autre !
	contextually correct:	T’inquiète, je **la** tuerai pour toi.
	incorrect:	T’inquiète, je **le** tuerai pour toi.
4	context:	Oh je déteste les papillons. Regarde, il y en a un autre !
	contextually correct:	T’inquiète, je **le** tuerai pour toi.
	incorrect:	T’inquiète, je **la** tuerai pour toi.

Blocks of 4 pairs for testing pronoun translation (Bawden 2018, p. 161)

Contrastive pairs (2)

- Pronoun set (Bawden et al. 2018)
  - 100 blocks, personal and possessives
  - generate 4 alternatives for the translation of each antecedent: (a) reference; (b) correct but opposite gender; (c, d) inaccurate (F/M)
  - contrastive pair: F/M pronoun
  - expected rankings: for (a) and (b) the correct gender; for (c) and (d) gender of the inaccurate translation (“contextually correct”)
  - ➢ results: the best system designed by Bawden et al. (2018) achieves 72.5% accuracy vs. 50% for non-contextual NMT

- ContraPRO (Müller et al., 2018)
  - 12,000 occ. of *it* from EN/DE Open Subtitles
  - possible translations by *er*, *sie* or *es* (4k each)
    - antecedents found automatically on both sides, with some confidence checks
    - most antecedents (58%) in previous sent.
  - wrong alternatives with random replacements
  - ➢ results: context-aware models (hierarchical) reach 64% (vs. 33% for a non-contextual baseline), especially when the antecedent is in the preceding sentence

Evaluating pronoun translation in subtitles

- Experiment with a contextual NMT system, window of 100 words
  - BLEU does not change with respect to a baseline NMT (34.9)
  - METEOR increases slightly from 0.60 to 0.62
  - manual inspection shows improvement of 2nd person pronouns
    - ➢ observed better handling of the politeness level (*tu/vous*) thanks to context
    - ➢ measurable with METEOR restricted to a list of words: increase from 0.54 to 0.66
      - *tu, toi, te, t', ton, ta, tes, vous, votre, vos*

❖Reference translations of subtitles are often disconcerting

Source	Baseline NMT (Transf.)	Contextual NMT	Reference
You don't actually believe that story, do you?	Tu ne crois pas vraiment à cette histoire ?	Vous ne croyez pas vraiment cette histoire, n'est-ce pas ?	Ne me dites pas que vous y croyez.
Besides, I'll owe you one. And I have every intention of collecting, ma'am.	En plus, je te revaudrai ça. Et j'ai l'intention de collecter, madame.	En plus, je vous en devrai une. Et j'ai l'intention de collecter, madame.	En outre, je vous serai redevable. Soyez sûre que je m'en souviendrai.
Major, I-- Look, I don't care what kind of wager you made with your pals. - Leave me alone. - Wager? I can take you anywhere you wanna go.	- Major, je... Je me fiche du pari que tu as fait avec tes copains. - Laisse-moi tranquille. - Wager ? Je peux t'emmener où tu veux.	- Major, je... Je me fiche du pari que vous avez fait avec vos amis. - Laissez-moi tranquille. - Wager ? Je peux vous emmener où vous voulez.	- Major, je... J'ignore quel type de pari vous avez fait avec vos amis. - Laissez-moi tranquille. - Un pari ? Je vous escorte où vous voudrez.

Evaluation of discourse structure and connectives

Exploratory metrics

- Theories of discourse structure (RST, SDRT, DTAG, CCR) are difficult to use
- Features related to discourse structure
  - Scarton and Specia (2015) defined a taxonomy for MT quality estimation
  - Lapshinova-Koltunski and Hardmeier (2017), Šoštarić et al. (2018) use contrastive linguistics at the discourse level: NMT outperforms SMT
  - automatic metrics involving discourse structure (sentence-level RST parse trees) correlate positively with human judgments of SMT (Joty et al., 2017)
- Relations conveyed explicitly by discourse connectives
  - Smith and Specia (2018) designed a discourse-aware metric that compares embeddings of source and target connectives, and was validated on legacy EN/FR SMT outputs (<2014)
  - ACT metric (Hajlaoui and Popescu-Belis 2013) showed that strategies for connective labeling do improve their translation by SMT (Meyer and Popescu-Belis, 2012; Meyer et al., 2015)

Quantitative evaluations of connectives

- *ACT: Accuracy of Connective Translation*
  - automatic count of correct connectives
  - uses automatic alignment to find out:
    - how *C* is translated in the reference
    - how *C* is translated in the candidate
  - compares the two translations of *C*
    - identical, “synonym”, incompatible, missing

- Results on 200 occurrences
  - within 5% of human ratings
  - can be improved by submitting litigious sentences (ca. 15%) to human judges
  - ➢ connective labeling can help PBSMT

- WMT 2019 EN/CZ shared task on discourse (Rysová et al. 2019)
  - topic-focus articulation and discourse connectives (including multi-word and alternative lexicalizations)
  - manual evaluation by linguists who compare MT output to English source

- Results on 100 documents
  - average of 80% agreement
  - 4 in-house systems (Transformers with or without context) and 1 online system
  - ➢ NMT quite on par with the reference

Conclusion

How should we measure document-level quality?

- Document-level quality = capacity to correctly translate discourse phenomena
  - e.g. cohesion, anaphora, coreference discourse relations / structure / connective

- Overall, discourse *divergencies* seem less frequent than lexical or syntactic ones
  - smaller potential for errors
  - often solved using local features
  - ➢ document-level evaluation is hard

1. Reference-based evaluation
   - BLEU/TER: OK in controlled experiments
   - metrics restricted to certain words (METEOR, APT, ACT): may capture only large variations
   - *The more human-like the translation, the less appropriate the reference-based metrics*
2. Contrastive sets: need probability estimates
3. Human annotators: still the final word
   - imperfect agreement, costly, not repeatable
   - test suites can accelerate the process

What have we learned about document-level quality?

- Many assessments of discourse quality of NMT, sometimes compared to SMT
- Often demonstrate some small benefits of context-aware NMT models
- But progress remains to be made
  - only unstructured representations of context are currently used (Kim et al. 2019)
  - minimal learning of anaphora resolution (Voita et al. 2018)
  - a lot of room for improvement on lexical cohesion (Bawden et al. 2018)

- Context-aware NMT models
  - concatenated sentences
    - work surprisingly well
  - multiple encoders
  - hierarchical networks

- What priority should be given to discourse?
  - must be solved for FAHQMT
  - plays a role in claims of human parity
  - infrequent divergencies, but potential very detrimental because they are difficult to spot by humans

My cat brought home a mouse that he hunted, and it was not dead but it was mortally wounded. What is the best way to kill it humanely?

[Google:] Mon chat a ramené à la maison une souris qu'elle a chassée. Elle n'était pas morte mais blessée à mort.
Quel est le meilleur moyen de le tuer humainement?

[DeepL] : Mon chat a ramené à la maison une souris qu'il chassait, et elle n'était pas morte, mais elle a été mortellement blessée. Quelle est la meilleure façon de le tuer humainement ?