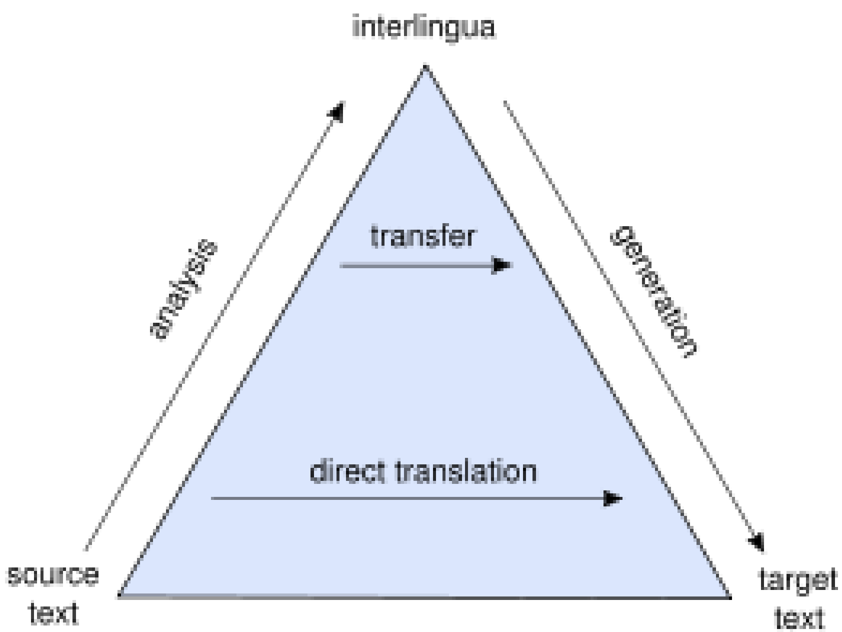
Document-level Statistical MT: from Connectives to Pronouns

Andrei Popescu-Belis
Idiap Research Institute, Martigny (VS)

"Machine Translation meets Translators"
Workshop at the University of Zurich, May 16, 2017

Since its start in the 1950s, and especially in the past 20 years, machine translation has made less and less use of linguistics



- State-of-the-art MT is slipping down the MT pyramid
- From rule-based, to example-based, to statistical systems
  - Within rule-based: from inter-lingua (representing meaning), to transfer (syntactic), to direct
- Neural MT: opaque interlingua?

The success of statistical MT

- "*Whenever I fire a linguist, our system performance improves*"
  - said Frederik Jelinek around 1980, marking the statistical turn in automatic speech recognition, followed later by machine translation
- What is statistical MT?
  - translation as a noisy channel (Weaver 1947, then Brown et al. 1993)
    1. Learn n-gram based translation and language models.
    2. Decode the source sentence: find the target sentence that maximizes the probabilities given by the translation model and the language model.
- Until recently, had state of the art performance
  - phrase-based or hierarchical SMT, or direct rule-based MT
  - since 2015, neural networks for MT have reached higher performance

Formal definition of SMT

- Goal: given *s*, find *t* which maximizes P(*t*|*s*)

- Rewritten using Bayes's theorem as:

$$\underset{t \in TL}{\operatorname{argmax}} P(t \mid s) = \underset{t \in TL}{\operatorname{argmax}} \left(P(s \mid t) \cdot P(t) \right)$$

translation model

language model

Formal definition of NMT

- Artificial neural networks: units/activation + connections/strengths
- How NMT works (Cho et al., EMNLP 2014)
  - represent words as individual units → learn to encode an abstract representation of a source sentence using stacked layers of units → decode representation into a foreign sentence
- Key additional contribution
  - "attention mechanism" (Bahdanau, Cho and Bengio, ICLR 2015)
- Enhancements to outperform SMT (Sennrich et al., WMT 2016)
  - character-based NMT for unknown words (byte-pair)
  - training on parallel data obtained from SMT output
  - very large computing power using GPUs (e.g., Google NMT)

Document-level machine translation

- Statistical or neural MT: efficient, good coverage, readable
- But systems always translate sentence by sentence
  - do not propagate information along a series of sentences
- Discourse information is helpful for coherent text translation
  - referring information, lexical chains: noun phrases, terms, pronouns
  - argumentative relations, as signaled by discourse connectives
  - verb tense, mode, aspect | style, register, politeness

Plan of this talk

1. Motivation and method

2. Document/discourse-level linguistic features for MT
   a. Disambiguation of English discourse connectives for MT
   b. Translation of English verb tenses into French
   c. Towards coherent translation of referring expressions
      i. coreference similarity as a criterion for MT from Spanish into English
      ii. consistent translation of repeated nouns from Chinese and German into English

3. Conclusion and perspectives

Credits

- Large collaboration started in 2010 supported by the Swiss National Science Foundation through two consecutive **Sinergia** projects

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

  COMTIS: Improving the coherence of MT by modeling inter-sentential relations

  MODERN: Modeling discourse entities and relations for coherent MT

  Also with support from the SUMMA EU project

- Research groups and people
  - **Idiap NLP group**: Thomas Meyer, Ngoc Quang Luong, Najeh Hajlaoui, Xiao Pu, Lesly Miculicich Werlen, Jeevanthi Liyanapathirana, Catherine Gasnier
  - **University of Geneva, Department of Linguistics:** Jacques Moeschler, Sandrine Zufferey, Bruno Cartoni, Cristina Grisot, Sharid Loaiciga
  - **University of Geneva, CLCL group:** Paola Merlo, James Henderson, Andrea Gesmundo
  - **University of Zurich, Institute of Computational Linguistics**: Martin Volk, Mark Fishel, Laura Mascarell, Annette Rios Gonzales, Don Tuggener
  - **Utrecht Institute of Linguistics**: Ted Sanders, J. Evers-Vermeul, Martin Groen, Jet Hoek

1. MOTIVATION AND METHOD

Examples: problems with discourse connectives

- Source: Why has no air quality test been done on this particular building ***since*** we were elected?
- SMT: Pourquoi aucun test de qualité de l' air a été réalisé dans ce bâtiment ***car*** nous avons été élus ?
- Human: Comment se fait-il qu'aucun test de qualité de l'air n'ait été réalisé dans ce bâtiment ***depuis*** notre élection?

- Source: What stands between them and a verdict is this doctrine that has been criticized ***since*** it was first issued.
- SMT: Ce qui se situe entre eux et un verdict est cette doctrine qui a été critiqué ***parce qu'*** il a d'abord été publié.
- Human: Seule cette doctrine critiquée ***depuis*** son introduction se trouve entre eux et un verdict.

Example: problems with verb tenses

- Source: Grandmother ***drank*** three cups of coffee a day.
- SMT: Grand-mère ***a bu*** trois tasses de café par jour.
- Human: Grand-maman ***buvait*** trois tasses de café par jour.

- Source: ... that we ***support*** a system that ***is*** clearer than the current one ...
- SMT: ... que nous ***soutenir*** un système qui ***est*** plus claire que le système actuel ...
- Human: ... que nous ***soutenons*** un système qui ***soit*** plus clair que le système actuel ...

Example: problem with NP coherence

- Source: Am 3. Juni schleppten Joe, Mac und ich die erste Traglast zum Lager II, während ***die Träger*** die unteren Lager mit Vorräten versorgten. [..] Am nächsten Morgen kamen ***die Träger*** unbegleitet vom Lager II zu uns herauf, als wir noch in den Schlafsäcken lagen.

- SMT: Le 3 Juin Joe, Mac, et j'ai traîné la première charge au camp II, tandis que ***le support*** fourni avec le roulement inferieur fournitures. [...] Le lendemain matin, ***le transporteur*** est arrive seul à partir de Camp II a nous, car nous étions encore dans leurs sacs de couchage.

- Human: Le 3, Joe, Mac et moi montâmes les premières charges au camp II, tandis que ***les porteurs*** faisaient la navette entre les camps inferieurs. [...] Nous étions encore dans nos sacs de couchage, le lendemain matin, lorsque ***les porteurs*** arrivèrent du camp II.

Examples: problems with pronouns

- Source: The table is made of wood. ***It*** is magnificent.
- SMT: La table est faite de bois. ***Il*** est magnifique.
- Human: La table est en bois. ***Elle*** est magnifique.

- Source: The European commission must make good these omissions as soon as possible. ***It*** must also cooperate with the Member States …
- SMT: La commission européenne doit réparer ces omissions dès que possible. ***Il*** doit également coopérer avec les états membres …
- Human: … ***Elle*** …

Summary of our goals

			1. Connective	2. Pronoun	3. Verb tense		
The matrix	*has been reduced*	*four times,*	*since*	*it*	*was*	*too large.*	
La matrice	*a été réduite*	*quatre fois,*	*depuis qu'*	*il*	*a été*	*trop grand.*	✘
			car	*elle*	*était*	*trop grande.*	✓

Current machine translation systems: red

Using longer-range dependencies: green

Method

1. Linguistic analyses
Cohesion markers for MT
Features for classification
Cross-linguistic perspective

2. Corpus data and annotation
Define tagset and guidelines
Locate problematic examples
Execute annotation and deliver data

3. Automatic labeling of cohesion markers
Build and test classifiers using surface features

4. SMT of labeled texts
Phrase-based SMT for labeled texts
Factored SMT models using labels

5. Evaluation
Define metrics of coherence
Performance of past systems
Apply metrics

Method

1. Define and analyze the phenomena to target
    - design theoretical models, keeping in mind objective and tractability
    - propose features for automatic recognizers
2. Create data for training and evaluation
    - define labeling instructions
    - annotate data sets (which can also be used for corpus linguistics)
    - validate linguistic models through empirical studies
3. Automatic disambiguation (= labeling = classification = recognition)
    - design and implement automatic classifiers
    - e.g. using machine learning over annotated data, based on surface features
4. Combine the automatically-assigned labels with MT
    - adapt MT systems (SMT or RBMT) or design new text-level translation models and decoding algorithms
5. Evaluation
    - assess improvements for the targeted phenomena and overall quality

Putting the method into application

- Phenomena discussed in this talk

  a. Discourse connectives   b. Verb tenses   c. Nouns/pronouns

- Languages

  English, French, German, Italian, Arabic, Chinese, Spanish

- Domains/corpora
  - parliamentary debates: Europarl (EU languages)
  - transcribed lectures: TED (ALL)
  - Alpine Club yearbooks: Text+Berg (FR, DE)
  - news: data from the Workshops on SMT (ALL)

2.a. DISAMBIGUATION OF ENGLISH DISCOURSE CONNECTIVES

What are discourse connectives?

- Small words, big effects
  - signal discourse relations between sentences or clauses
  - additional, temporal, causal, conditional, etc.

- Theoretical descriptions
  - Rhetorical Structure Theory (Mann and Thompson)
  - Discourse Representation Theory (Asher et al.)
  - Cognitive approach to Coherence Relations (Sanders et al.)
  - annotation-oriented: Penn Discourse Treebank (PDTB) (Prasad, Webber, Joshi et al.)

- Connectives are challenging for translation because they may convey different relations, which are translated differently
  - *while* contrastive or temporal: French *mais* or *pendant que*
  - *since* causal or temporal: French *puisque* or *depuis que*

- Wrong translations of connectives lead to:
  - low coherence or readability
  - distorted relationships between sentences
  - correct relations are sometimes impossible to recover

Annotation of discourse connectives for translation (Cartoni, Meyer, Zufferey)

- Penn Discourse Tree Bank (PDTB): complex hierarchy of senses
  - difficult to annotate, not necessarily relevant to MT
- Annotation through translation spotting
  - annotators identify the human translation of each connective (in Europarl)
  - observed translations are clustered into *a posteriori* "senses" relevant to MT
  - fewer labels, cheaper to annotate (e.g. *while* has 21 PDTB labels vs. 5 here)

Connective	Training set EP	Training set PDTB	Training set Distribution of labels (%)	Testing set EP	Testing set PDTB	Testing set Distribution of labels (%)
although	168	312	Ct: 68.9; Cs: 31.1	15	16	Ct: 48.4; Cs: 51.6
however	348	450	Ct: 47.8; Cs: 52.2	70	35	Ct: 47.6; Cs: 52.4
meanwhile	102	177	Ct: 77.3; T: 22.7	28	14	Ct: 76.2; T: 23.8
since	339	174	Ca: 38.7; T: 59.6; T/Ca: 1.7	82	10	Ca: 30.4; T: 67.4; T/Ca: 2.2
(even) though	276	306	Ct: 33.3; Cs: 66.7	69	14	Ct: 33.7; Cs: 66.3
while	236	744	Ct: 14; Cs: 23; T: 15; T/Ct: 46.6; T/Ca: 1.4	58	37	Ct: 22.8; Cs: 33.7; T: 9.8; T/Ct:
yet	326	99	Ct: 23.2; Cs: 29.8; Adv: 47	77	2	Ct: 30.4; Cs: 19; Adv: 50.6
Total	1795	2262	–	399	128	–

Features for the automatic disambiguation of connectives

Hong Kong-NNP trade figures illustrate-PRESENT the toy makers' reliance on factories across the border-NN. -JOINT- In-IN 1989's first seven months, -JOINT- domestic exports fell-VBD-PAST-1 29%, to HK$3.87 billion-NN, -CONTRAST- while-IN re-exports-NN rose-VBD-PAST 56%, to HK$11.28 billion-NN.

- syntactic features
  - connective, punctuation, context words, context tree structures, auxiliary verbs
- WordNet antonymy features
  - similarity scores (word distance) and antonyms from the clauses
- TimeML features
- discourse relation features
  - discourse relations from a discourse parser
- polarity features
  - using a polarity lexicon, count positive and negative words, account for negation
- translational features
  - baseline translation (e.g. *tandis que*), sense from dictionary (*contrast*), position (*25)*

- Extracted from the current and the previous sentences

Automatic labeling of connectives
(Th. Meyer)

- For each (new, unseen) discourse connective
  - given the features extracted from the text
  - determine its most probable label ("sense")

- Use of machine learning for classification
  - Maximum Entropy classifier

  1. trained on manually labeled data
     - experimented with PDTB and/or Europarl
  2. tested on unseen data

Automatic connective labeling: F1 scores

Data	Method	although	however	meanwhile	since	(even) though	while	yet
Training (c.v.)	`All_Features`	0.69 ± 0.04	0.85 ± 0.05	0.86 ± 0.01	0.93 ± 0.05	0.77 ± 0.04	0.76 ± 0.04	0.88 ± 0.07
Test: Europarl and PDTB (WSJ s. 23)	`Majority class`	0.52	0.52	0.76	0.68	0.66	0.34	0.51
	`All_Features`	0.58	**0.73**	0.71	**0.90**	0.69	0.45	**0.78**
	`Best`	0.61	0.60	0.74	0.87	**0.71**	0.43	0.72
	`All_Synt+Dep`	**0.65**	0.67	**0.79**	0.89	0.7	**0.47**	0.72
Test: Europarl	`All_Features`	0.60	**0.69**	0.79	**0.90**	0.67	0.45	**0.78**
	`Best`	**0.80**	0.56	0.82	0.85	**0.72**	0.43	0.74
	`All_Synt+Dep`	0.73	0.66	**0.89**	0.88	0.71	**0.50**	0.73
Test: PDTB (WSJ s. 23)	`All_Features`	**0.56**	**0.83**	0.57	0.90	**0.79**	**0.46**	**1.0**
	`Best`	0.44	0.69	0.57	**1.0**	0.64	0.43	0.0
	`All_Synt+Dep`	**0.56**	0.69	0.57	**1.0**	0.64	0.43	0.50

- Findings
  - scores compare well to human agreement levels (80-90%)
  - classifying each connective separately is better than jointly
  - using all features is the best option

How do we use labeled connectives in SMT?

Four possible methods have been tested

1. Replace in the system's phrase table all unambiguous occurrences of the connective with the correct one

2. Train the system on (a) manually or on (b) automatically labeled data, with labels concatenated to words (e.g., *while_Temporal*)

3. Use a connective-specific SMT system only when the connective labeler is confident enough (otherwise use a baseline one)

4. Use Factored Models as implemented in the Moses system
   – word-level linguistic labels are separate translation features
   – a model of labels is learned when training, then used when decoding

How do we measure the improvement of connective translation? (Meyer, Hajlaoui)

- Measuring translation quality
  - subjective measures: fluency, fidelity → too expensive for everyday use
  - objective, reference-based measures: BLEU (or METEOR, etc.)
    - comparison of a candidate text with one or more reference translations in terms of common n-grams (usually from 1 to 4)
  - connectives are not frequent → small effects expected on BLEU scores

- Count how many connectives are correctly translated:
  ACT metric [Accuracy of Connective Translation]
  - given a source sentence with a discourse connective *C*
  - use automatic alignment to find out:
    - how *C* is translated in the reference and in the candidate translations
  - compare the translations: identical | "synonymous" | incompatible | absent

Improvement of SMT and connectives

1. Modified phrase table
   Tested on ~10,000 occurrences of 5 types: **34%** improved, **20%** degraded, **46%** unchanged

2. Concatenated labels
   (a) trained on manually labeled data: **26%** improved, **8%** degraded, **66%** unchanged
   (b) trained on automatically labeled data: **18%** improved, **14%** degraded, **68%** unchanged

3. Thresholding based on automatic labeler's confidence
   With two connectives only: improvement of **0.2**-**0.4** BLEU points

4. **Factored models in Moses SMT**

Languages	Test set	System	BLEU	Δ	p	ACT	Δ	p
EN/FR	nt2012	baseline	26.1			56.28		
		labeled connectives	25.8	-0.3	**	57.68	1.40	*
	nt2010	baseline	24.4			68.12		
		labeled connectives	24.3	-0.1	**	68.60	0.48	*
	nt2008+sy2009	baseline	28.9			61.36		
		labeled connectives	29.2	0.3	*	60.94	-0.42	*
EN/DE	nt2012	baseline	11.8			62.28		
		labeled connectives	11.8	0.0	n/s	65.08	2.80	**
	nt2010	baseline	15.0			62.42		
		labeled connectives	15.0	0.0	n/s	69.28	6.86	***
	nt2008+sy2009	baseline	13.0			71.06		
		labeled connectives	13.1	0.1	n/s	70.30	-0.76	n/s

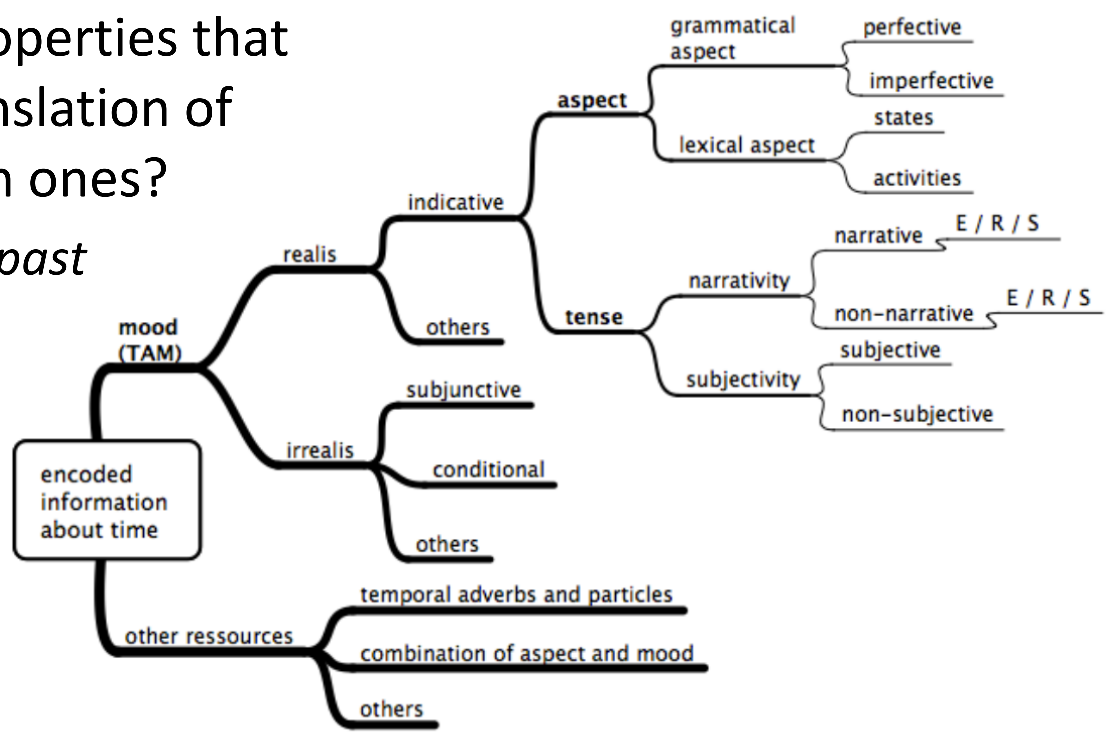
2.b. TRANSLATING VERB TENSES

Cross-lingual modeling of verb tenses
(Grisot and Moeschler)

- Two well-known models
  - event time, reference time, speech time (Reichenbach)
  - four classes of aspect (Vendler)
- What are the relevant properties that would enable correct translation of English tenses into French ones?
  - focus on English *simple past*
- Theoretical hypothesis:

  simple past narrative
  → *passé simple* or *passé composé*

  simple past non-narrative
  → *imparfait*



Empirical studies of tense translation

- Approaches: narrativity-based vs. general tense correlation

1. Annotation of narrativity (C. Grisot)
   - English/French parallel corpus
   - 576 EN simple past verb phrases
   - inter-annotator agreement on 71% of instances: $\kappa = 0.44$

   ➔ narrativity correctly predicts 80% of translated tenses

2. Annotation of translated tense for all English VPs (S. Loaiciga)
   - rules for precise alignment of VPs in Europarl
   - annotated ca. 320,000 VPs, with about 90% precision

   ➔ confirmed divergencies between EN and FR tenses

Observed EN/FR tense divergencies for 322,086 verb phrases (Loaiciga)

French	English: Past continuous	Past perfect continuous	Past perfect	Present continuous	Present perfect continuous	Present perfect	Present	Simple past	Total
Imparfait	462 54%	7 27%	**365** **24%**	146 1%	18 2%	463 1%	1 510 1%	**8 060** **21%**	11 031 3%
Impératif				37 0%	1 0%	6 0%	203 0%	11 0%	258 0%
Passé composé	139 16%	2 8%	**214** **14%**	282 1%	325 33%	**26 521** **61%**	1253 1%	**19 402** **49%**	48 138 15%
Passé récent			1 0%	8 0%	3 0%	187 0%	2 0%	3 0%	204 0%
Passé simple	4 1%		6 0%	16 0%	2 0%	54 0%	42 0%	374 1%	498 0%
Plus-que-parfait	27 3%	8 31%	**782** **52%**	2 0%	4 0%	217 1%	22 0%	1 128 3%	2 190 1%
Présent	216 25%	9 35%	102 7%	18 077 96%	617 63%	**14 736** **34%**	211 334 97%	**9 779** **25%**	254 870 79%
Subjonctif	15 2%		28 2%	258 1%	6 1%	**1 053** **2%**	2 969 1%	568 1%	4 897 2%
Total	863 100%	26 100%	1 498 100%	18 826 100%	976 100%	43 237 100%	217 335 100%	39 325 100%	322 086 100%

Features for automatic prediction of narrativity or (directly) translated tense

If the situation were-VBD-PAST-SV-1-0 to change-VBP-INFINITIVE-0-0-0, it would-MD-CONDITIONAL-CSV-0-0 clearly also change-VBP-INFINITIVE-0-0-0 as far as-synch we are-VBP-PRESENT-CSV-0-0 concerned-VBN-0-0-0-0.

- all verbs in the current and previous sentences
- word positions
- verb POS and trees
- auxiliaries and tenses
- TimeML features
- temporal connectives (from a hand-crafted list)
- synchrony/asynchrony of the connectives
- semantic roles
- imparfait indicator: yes/no
- subjonctif indicator: yes/no

- Extracted from the current and the previous sentences

Automatic annotation: results

- Using a maximum entropy classifier

1. Automatic annotation of narrativity (+/-)
   – training on 458 instances, testing on 118
2. Prediction of translated tense
   – training/testing on 196’000 instances with 10-fold cross-validation

Improvements of SMT using narrativity

- Scores from human evaluators
  1. Is the narrativity label correct?
  2. Are verb tenses and lexical choices improved?

Criterion	**Rating**	**N.**	**%**	Δ
Labeling	correct	147	71.0	
	incorrect	60	29.0	
Verb tense	+	35	17.0	+9.7
	=	157	75.8	
	−	15	7.2	
Lexical choice	+	19	9.2	+3.4
	=	176	85.0	
	−	12	5.8	

Improvements of SMT using predicted tense labels

- Oracle = prefect prediction
- BLEU scores per target tense

	Baseline	Oracle	Predicted	# Sent.
Imparfait	24.10	25.32	24.57	122
Passé composé	29.80	30.82	30.08	359
Impératif	19.08	19.72	18.70	4
Passé simple	13.34	16.15	14.09	6
Plus-que-parfait	21.27	23.44	23.22	17
Présent	27.55	27.97	27.59	2,618
Subjonctif	26.81	27.72	26.07	78
Passé recent	24.54	30.50	30.08	3

- Manual evaluation of a sample

French tense	System	TAM Incorrect	TAM Correct $\neq$ ref	TAM Correct = ref
Imparfait	Baseline	**82**	**15**	**41**
	Predicted	**42**	**23**	**73**
	Oracle	**13**	**4**	**121**

2.c. REFERENTIAL COHERENCE IN MT

Can we improve MT of nouns using document/discourse-level information?

1. Translate nouns so as coreference relations from the source text are preserved in the translated text
   – challenge: compute coreference automatically

2. Translate repeated nouns consistently, i.e. using the same translation
   – challenge: learn when to enforce consistency

Previous work on consistency and coreference

- How do human and MT consistency compare? Is consistency correct?
  - it is often the case that there is "one translation per discourse" (Carpuat 2009)
  - "the trouble with MT consistency"(Carpuat and Simard, 2012)
    - systems are often (and wrongly) more consistent than humans, due to lack of coverage
    - inconsistencies (i.e. errors) are often due to semantic/syntactic mistakes
  - human translators are often more consistent with nouns than verbs (Guillou 2013)
  - encourage consistent translation by "caching" (Tiedemann, 2010; Gong et al., 2011)

- How can coreference help MT?
  - anaphora resolution is somewhat helpful for pronoun translation, but surface features do better (Hardmeier et al. 2015; Guillou et al. 2016; Loaiciga et al. *in preparation*)

- Coreference is a good reason to enforce noun consistency, but surface features can also help to decide when/how to correct inconsistencies

2.c.i. USING A COREFERENCE SCORE TO RE-RANK MT HYPOTHESES

Using coreference similarity for MT

- Principle
  - preserve the information conveyed in translation: here, information about the entities (i.e. grouping of mentions)

  → *better translations should have coreference links that are more similar to those of the source text*

- Maximize a global coreference similarity score by re-ranking hypotheses from the Moses SMT decoder
  - Spanish-to-English translation using gold coreference links on the source side, from AnCora-ES (Recasens and Martí 2010), as test data

Miculicich Werlen L. & and Popescu-Belis A. (2017) - Using Coreference Links to Improve Spanish-to-English Machine Translation. *Proceedings of the EACL Workshop on Coreference Resolution Beyond OntoNotes (CORBON)*, Valencia, p. 30-40, 4 April 2017.

Motivating example

Source	Human Translation	Baseline SMT
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.	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.	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]$_{c1}$** songs.
Pudiera considerarse un viaje fallido, porque **[Ø]$_{c1}$** no encuentra **[su]$_{c1}$** objetivo, pero el azar **[le]$_{c1}$** conduce a una pequeña comunidad...	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...	It could be considered **[a failed trip]$_{c3}$** because **[it]$_{c3}$** does not find **[its]$_{c3}$** objective, but the chance leads Ø to a small community...

Challenge: compute a reliable "coreference score" for a translation

- For any candidate translation, measure the similarity between its coreference links and those of the source text

1. Apply a coreference resolver to the source text and the translation
   - NB: this is the major source of errors in estimating the CSS
   - NB: in this work, we use ground truth links on the source side (fixed), and only run automatic coreference resolution (Stanford Core NLP Tools) on translations

2. Project mentions from the candidate translation back to the source (i.e. referring expressions: nouns, pronouns)

3. Apply existing metrics for evaluating coreference links on the source text
   - 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

   ➔ CSS (coreference similarity score): average of MUC, B3 and CEAF

Empirical verification: CSS increases with better translations

(on 3k words from AnCora-ES)

Hypothesized Translation 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*

Automatic Coreference Quality

F1 scores (%)

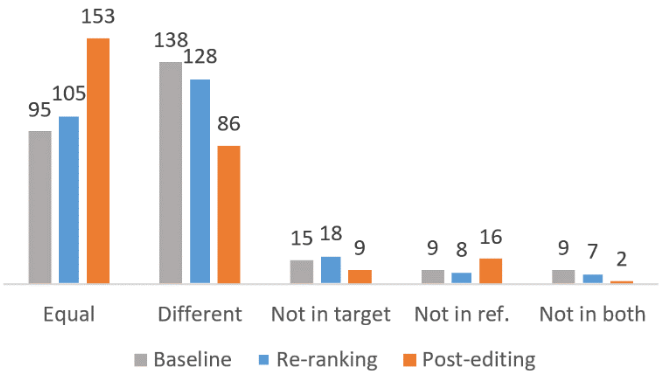
Using the CSS for document-level MT

- Phrase-based ES-EN statistical MT: Moses
  - trained on WMT 2013 (14M sentences)
  - tuned on News Commentary 2011 (5.5k s.)
  - tested on News Test 2013 (3k s., BLEU = 30.8)
- For each sentence of a translated text
  - get from Moses the 1000-best hypotheses
  - select those that differ in the translations of mentions
- Beam search to maximize the CSS
  - starting from the first sentence, search among the hypotheses for those that improve the text-level CSS

Evaluation
(10 test documents, with our translations)

Metric	PBSMT	NMT	PBSMT + Re-ranking
BLEU	46.5±4.3	46.9±3.7	41.7±3.9
Accuracy of pronoun translation	0.35±0.07	0.37±0.07	0.40±0.1
Accuracy of noun translation	0.78±0.08	0.78±0.07	0.74±0.01

- The number of pronouns identical to the reference translation increases
  - especially for a second approach, based on post-editing mentions
    - *see (Miculicich & APB, 2017)*



Findings

- The principle of "maximizing coreference similarity with the source" fails to increase the accuracy of noun translation
  - possible causes
    - imperfect (ca. 60-70%) automatic coreference resolution (→ no simple solution)
    - imperfect use of the criterion in SMT (→ could try Docent)
    - optimal translation is not among 1000-best hypotheses (20% of the cases)
  - requires coreference resolution for every translation hypothesis
- Our 2nd method has promising results for pronoun translation: post-editing the mentions & maximizing coreference features

→ Narrow our focus to repeated nouns

- partial overlap with coreference, but more tractable

2.c.ii. ENFORCING TRANSLATION CONSISTENCY OF REPEATED NOUNS

First attempt: consistent translation of noun compounds (DE, ZH → EN)

- Motivating example

  Src: das Bundesamt für Landestopographie […] **dieses Amt** war in der Lage,

  Ref: Seul **cet office** était en mesure,

  SMT: Que **ce poste** était dans la situation,

- Assumptions: given a compound (*XY*) and a subsequent occ. of the head noun (*Y*)
  - assume that the latter is a mention of the former (co-reference)
  - assume the translation of *Y* in *XY* is more accurate than of *Y* alone
- Method: replace the translation of the second occurrence with the first one
- Challenges
  - avoid non-compound XY, and non-coreferent XY/Y pairs
  - correctly identify the translations of XY and Y

Mascarell L., Fishel M., Korchagina N., and Volk M. (2014) - Enforcing consistent translation of German compound coreferences. In Proceedings of the 12th Konvens Conference, Hildesheim, Germany.

Pu X., Mascarell L., Popescu-Belis A., Fishel M., Luong N.Q., & Volk M. (2015) - Leveraging Compounds to Improve Noun Phrase Translation from Chinese and German. ACL-IJCNLP 2015 Student Research Workshop, Beijing, p.8-15.

Example of a Chinese compound

1. Chinese source sentence	她以为自买了双两英寸的高跟鞋， 但实际上那是一双三英寸高的鞋。
2. Segmentation, POS tagging, identification of compounds and their co-reference	她#PN 以为#VV 自#AD 买#VV 了#AS 双#CD 两#CD 英寸#NN 的#DEG **高跟鞋#NN** ，#PU 但#AD 实际上#AD 那#PN 是#VC 一#CD 双#M 三#CD 英寸#NN 高#VA 的#DEC **鞋#NN** 。#PU
3. Baseline translation into English (statistical MT)	She thought since bought a pair of two inches high **heel**, but in fact it was a pair of three inches high **shoes**.
4. Automatic post-editing of the baseline translation using compounds	She thought since bought a pair of two inches high **heel**, but in fact it was a pair of three inches high **heel**.
5. Comparison with a human reference translation	She thought she'd gotten a two-inch heel but she'd actually bought a three-inch **heel**. ✓

Improvement of SMT using compounds

- Test data for SMT: ZH/EN and DE/FR
  - training sets: about 200k sentences | tuning: about 2k sentences
  - testing: 800/500 sentences with ca. 250 *XY/Y* pairs

- BLEU scores
  - ZH/EN: 11.18 → 11.27 | DE/FR: 27.65 → 27.48

- Comparison of the *Y* translations (in % of total)
  - our 2 systems are closer to the reference than the baseline

			CACHING		POST-EDITING	
			= ref	≠ ref	= ref	≠ ref
ZH/EN	BASELINE	= ref	59.3	*4.1*	42.3	*4.5*
		≠ ref	**13.8**	22.8	**20.3**	32.9
DE/FR	BASELINE	= ref	70.1	*10.3*	73.9	*5.0*
		≠ ref	**4.3**	15.2	**3.5**	17.5

Second attempt: consistent translation of repeated nouns

- Automatically enforcing consistent noun translations
  - learn whether two occurrences of the same noun must be translated identically or not, based on several features, but *not coreference*

- Method
  1. Detect two close occurrences of the same noun in the source
  2. Find their baseline translations by a PBSMT using word alignment
  3. If they differ, decide whether/how to edit: $1^{st} \rightarrow 2^{nd}$, or vice-versa
  4. Based on this decision, post-edit and/or re-rank the PBSMT output

Pu X., Mascarell L. & Popescu-Belis A. (2017) - Consistent Translation of Repeated Nouns using Syntactic and Semantic Cues. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Valencia, 5-7 April 2017.

Example

- *Source*: nach einfuehrung dieser **politik** […] die **politik** auf dem gebiet der informationstechnik […]

- *Reference:* once the **policy** is implemented […] the information technology **policy** […]

- *MT:* after introduction of **policy** […] the **politics** in the area of information technology[…]

Example

- *Source*: 赞扬 联合国 人权 事务 高级 专员 办事处 高度 优先 从事 有关 国家 机构 的 工作 , [...], 鼓励 高级 专员 确保 作出 适当 安排 和 提供 预算 资源

- *Reference:* commends the high priority given by the office of the united nations high commissioner for human rights to work on national institutions, [...] , encourages the high commissioner to ensure that appropriate arrangements are made and budgetary resources provided.

- *MT:* praise the human rights high commissioner was the high priority to offices in the country, [...] , to encourage senior **specialists** to make sure that make appropriate and provided budget resources.

Data and classifiers

- Training data = with the correct consistency decisions
  - source text + baseline MT output + reference translation
  - detect pairs of repeated source nouns, inconsistently translated by baseline
  - use word-aligned reference translation to set the correct decision
    - `if the two reference translations differ, then label as 'none' [else…]`
    - `if the reference translation is equal to one of the baseline translations, then paste this word over the other one ('1→2' or '2→1') [else…]`
    - `label as 'none'`
- Testing data = same as above (to test the classifier)
  or parallel (to test end-to-end MT)
- Extracted pairs (UN Corpora)
  - ZH/EN: 3,301 train, 647 test | DE/EN: 11,289 train, 695 test
- Classifiers
  - experimented with decisions trees, random forests, Naïve Bayes, SVM
  - syntactic and semantic features

Syntactic features

Features	Values
Source noun (Chinese)	专员
Distance in sentences between the two source occurrences	0
Translation of the first occurrence (labeled NN)	commissioner
Translation of the second occurrence (labeled NN)	specialists
Number of sibling nodes of the 1st occurrence	4
Number of sibling nodes of the 2nd occurrence	2
Sign of the difference between the above (+1, 0, −1)	1
Number of words of the 1st occurrence and its siblings	2
Number of words of the 2nd occurrence and its siblings	1
Sign of the difference between the above (+1, 0, −1)	1
Number of nodes in the first NP ancestor of 1st occ.	15
Number of nodes in the first NP ancestor of 2nd occ.	7
Sign of the difference between the above (+1, 0, −1)	1
Number of words in the first NP ancestor of the 1st occ.	6
Number of words in the first NP ancestor of the 2nd occ.	2
Sign of the difference between the above (+1, 0, −1)	1
Distance between the first NP ancestor and the 1st occ.	3
Distance between the first NP ancestor and the 2nd occ.	3
Sign of the difference between the above (+1, 0, −1)	0
Class (1, 2, 0)	1

```
(CP
     (IP
      (NP
       (NP (NR 联合国) (NN 人权) (NN 事务))
       (ADJP (JJ 高级))
       (NP (NN 专员) (NN 办事处)))
      (VP
       (ADVP (AD 高度))
       (VP
        (VP
         (ADVP (AD 优先))
         (VP (VV 从事)
          (NP
           (DNP
            (NP
             (ADJP (JJ 有关))
             (NP (NN 国家) (NN 机构)))
            (DEG 的))
           (NP (NN 工作)))))
        (PU ，)
```

```
(CC 并且)
        (VP
         (PP (P 鉴于)
          (NP
           (DNP
            (NP
             (ADJP (JJ 有关))
             (NP (NN 国家) (NN 机构)))
            (DEG 的))
           (NP (NN 活动))))
         (VP (VV 有所)
          (VP (VV 增加))))
        (PU ，)
        (VP (VV 鼓励)
         (NP
          (ADJP (JJ 高级))
          (NP (NN 专员)))
         (IP
          (VP (VV 确保)
           (VP
            (VP (VV 作出)
             (NP
              (ADJP (JJ 适当))
              (NP (NN 安排))))
            (CC 和)
            (VP (VV 提供)
             (NP (NN 预算) (NN 资源))))))
        (PU ，)
```

Semantic features

- For each of the two occurrences (1st and 2nd)
- Features of the *local context* (in source and target)
  - values of 3 surrounding words to the left and right, within the same sentence
- Features of the *discourse context* (in target only)
  - cosine similarity between the vector representation (word2vec) of the translated word and the vector of its context
    - context = average of 20 words before and 20 after the word
  - *interpretation*: if inconsistency is due to the sense ambiguity of the source noun, use semantic similarity to decide which of the two translations best matches its context

Data

UN data to train/test the classifiers					
Training			Testing		
Sent.	Words	Nouns	Sent.	Words	Nouns
150K	4.5M	11,289	7,771	225K	695
185K	3,4M	3,301	3,000	121K	647

WIT[3] data for building SMT						
	Training		Tuning		LM	
	Sent.	Words	Sent.	Words	Sent.	Words
DE-EH	193K	3.6M	2,052	40K	217K	4,4M
ZH-EN	185K	3,4M	2,457	54K	4,8M	800M

Noun pair classification, for ZH/EN and DE/EN, with 10-fold cross-validation

Prediction of correct translation for repeated nouns in ***Chinese***						
	Syntactic features		Semantic features		All features	
	Acc. (%)	κ	Acc. (%)	κ	Acc. (%)	κ
SVM	72.1	0.48	60.2	0.00	60.2	0.00
J48	74.5	0.54	60.2	0.00	73.9	0.51
RF	75.3	0.54	68.4	0.29	70.7	0.35
MaxEnt	76.7	0.65	69.5	0.32	**83.3**	**0.75**

Prediction of correct translation for repeated nouns in ***German***						
	Syntactic features		Semantic features		All features	
	Acc. (%)	κ	Acc. (%)	κ	Acc. (%)	κ
SVM	77.9	0.67	38.1	0.00	38.1	0.00
J48	77.0	0.66	64.8	0.45	79.7	0.69
RF	82.0	0.73	73.5	0.60	84.5	0.77
MaxEnt	80.8	0.71	76.8	0.65	**83.4**	**0.75**

Integration with MT

1. Post-editing
   – edit the baseline translation depending on the classifier's decision

2. Re-ranking
   – obtain the 10,000-best translation hypotheses from the SMT system
   – search among them for highest ranking one in which the repeated word is translated as predicted by the classifier
   – if none is found, keep the best hypothesis

3. Re-ranking + Post-editing
   – same as (2), but if none is found, post-edit the baseline translation

Classification and MT results (BLEU scores) for ZH/EN and DE/EN

	Syntactic features					Semantic features					All features				
	Acc.	κ	BLEU			Acc.	κ	BLEU			Acc.	κ	BLEU		
			PE	RR	RR+PE			PE	RR	RR+PE			PE	RR	RR+PE
Baseline	-	-	11.07	11.07	11.07	-	-	11.07	11.07	11.07	-	-	11.07	11.07	11.07
J48	66.3	0.42	11.17	11.20	11.30	33.1	0.00	11.07	11.07	11.07	33.1	0.00	11.07	11.07	11.07
SVM	71.9	0.53	11.23	11.27	11.33	33.1	0.00	11.07	11.07	11.07	62.1	0.43	11.18	11.26	11.26
RF	71.7	0.53	11.22	11.24	11.27	55.2	0.33	11.04	11.07	11.12	54.9	0.32	11.16	11.20	11.24
MaxEnt	73.7	0.60	11.27	11.33	**11.35**	56.1	0.34	10.87	11.11	11.18	72.5	0.56	11.21	11.33	**11.36**
Oracle	100	1.00	11.40	11.52	11.64	100	1.00	11.40	11.52	11.64	100	1.00	11.40	11.52	11.64

Table 4: Prediction of the correct translation (accuracy (%) and *kappa*) and translation quality (BLEU) for repeated nouns on the *Chinese test set*. Maximum Entropy was the best method found on the dev set.

	Syntactic features					Semantic features					All features				
	Acc.	κ	BLEU			Acc.	κ	BLEU			Acc.	κ	BLEU		
			PE	RR	RR+PE			PE	RR	RR+PE			PE	RR	RR+PE
Baseline	-	-	17.10	17.10	17.10	-	-	17.10	17.10	17.10	-	-	17.10	17.10	17.10
SVM	71.4	0.57	17.59	17.65	17.72	32.8	0.00	17.10	17.10	17.10	32.8	0.00	17.10	17.10	17.10
J48	70.5	0.56	17.59	17.61	17.70	48.2	0.23	17.13	17.27	17.33	69.4	0.54	17.56	17.60	17.66
RF	70.2	0.55	17.55	17.62	17.68	54.4	0.32	17.21	17.34	17.37	67.6	0.52	17.53	17.57	17.63
MaxEnt	78.3	0.67	17.63	17.66	**17.75**	63.5	0.49	17.39	17.47	17,49	68.7	0.53	17.58	17.59	**17.67**
Oracle	100	1.00	17.78	17.83	17.99	100	1.00	17.78	17.83	17.99	100	1.00	17.78	17.83	17.99

Table 5: Prediction of the correct translation (accuracy (%) and *kappa*) and translation quality (BLEU) for repeated nouns on the *German test set*. Maximum Entropy was the best method found on the dev set.

Pronoun MT: coreference (anaphora) or not?

- Active research topic, shared tasks since 2015
  - focusing on divergencies such as *it* → *il* | *elle* | *ce* | …
- Studies by Idiap's NLP group (Luong et al., 2016-7)
  1. Pronoun-aware language model
     - post-edit translated pronouns based on neighboring nouns
  2. Anaphora-aware decoder with uncertainty modeling
     - learn probabilities for pronoun translation based on probability distributions of the antecedents
- Many other studies
  - surface features outperform anaphora resolution
  - no need for antecedent, just a guess of translation

3. CONCLUSION AND PERSPECTIVES

Conclusion

- Long-range dependencies can be modeled thanks to linguistic theories, and their automatic annotation, although imperfect, can benefit SMT

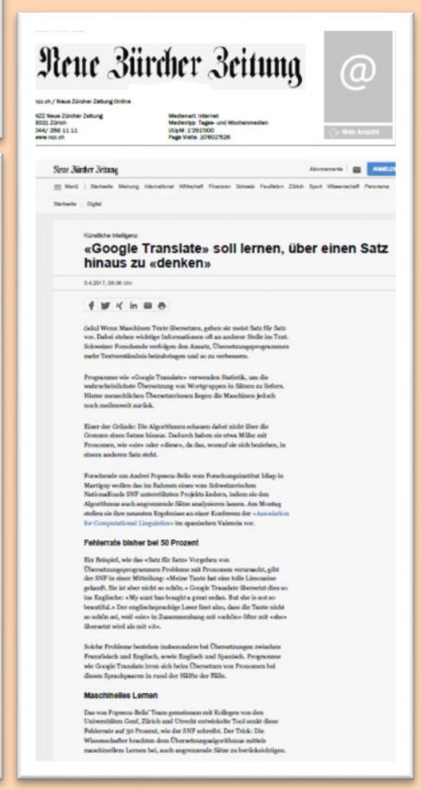
- Genuine collaboration between: theoretical linguistics and pragmatics, corpus linguistics, natural language processing, and machine translation

- Some outputs
  - publications: available from COMTIS and MODERN websites
  - resources: annotations of discourse connectives and verb phrases
  - software: automatic connective labeler, ACT and APT metrics

Perspectives

- Correct and consistent [pro]noun translation remains an open problem
  - improved anaphora/coreference resolution is beneficial to MT
  - but using only coreference-related *features* seems the best approach
  - dilemma: invest research in the classifiers or in the MT?
- Future work
  - word sense disambiguation and MT (especially for nouns)
  - larger use of context in neural MT (for nouns and pronouns)
  - how do we integrate these complex, heterogeneous knowledge sources into efficient and robust SMT or NMT systems?
- Sinergia MODERN and COMTIS: established discourse-level MT
  - worked on connectives and verb tenses, before pronouns/nouns
  - workshops every two years: DiscoMT 2013, 2015, 2017
  - shared tasks on pronoun prediction in translations: 2015, 2016, 2017

SNSF Press release on April 3, 2017 and subsequent press articles



THANK YOU FOR YOUR ATTENTION!
ANY QUESTIONS?

References

- Luong N.Q. & Popescu-Belis A. (2016) - A Contextual Language Model to Improve Machine Translation of Pronouns by Re-ranking Translation Hypotheses. *Proceedings of EAMT 2016 (19th Annual Conference of the European Association for Machine Translation)*, Riga, Latvia, special issue of the *Baltic Journal of Modern Computing*, vol. 4, n. 2, p. 292-304.
- Luong N.Q. & Popescu-Belis A. (2016) - Improving Pronoun Translation by Modeling Coreference Uncertainty. *Proceedings of WMT 2016 (First Conference on Machine Translation), Research Papers*, Berlin, Germany, p. 12-20.
- Luong N.Q., Popescu-Belis A., Rios Gonzales A., & Tuggener D. (2017) - Machine translation of Spanish personal and possessive pronouns using anaphora probabilities. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Valencia, 5-7 April 2017.
- Pu X., Mascarell L., Popescu-Belis A., Fishel M., Luong N.Q., & Volk M. (2015) - Leveraging Compounds to Improve Noun Phrase Translation from Chinese and German. *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, Beijing, p.8-15.
- Pu X., Mascarell L. & Popescu-Belis A. (2017) - Consistent Translation of Repeated Nouns using Syntactic and Semantic Cues. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Valencia, p. .
- Miculicich Werlen L. & and Popescu-Belis A. (2017) - Using Coreference Links to Improve Spanish-to-English Machine Translation. *Proceedings of the EACL Workshop on Coreference Resolution Beyond OntoNotes (CORBON)*, Valencia, 4 April 2017.
- Meyer T., Hajlaoui N., & Popescu-Belis A. (2015) - Disambiguating Discourse Connectives for Statistical Machine Translation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(7):1184-1197.

References (continued)

- Grisot C. & Meyer T. (2014) - Cross-Linguistic Annotation of Narrativity for English/French Verb Tense Disambiguation. *Proceedings of LREC 2014 (9th Int. Conf. on Language Resources and Evaluation)*, Reykjavik.
- Loaiciga S., Meyer T. & Popescu-Belis A. (2014) - English-French Verb Phrase Alignment in Europarl for Tense Translation Modeling. *Proceedings of LREC 2014 (9th Int. Conf. on Language Resources and Evaluation)*, Reykjavik.
- Mascarell L., Fishel M., Korchagina N., & Volk M (2014) - Enforcing Consistent Translation of German Compound Coreferences. *Proceedings of KONVENS 2014 (12th German Conference on Natural Language Processing)*, Hildesheim, Germany.
- Cartoni B., Zufferey S., Meyer T. (2013) - Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *Dialogue & Discourse*, 4(2):65-86.
- Zufferey S. & Cartoni B. (2012) - English and French causal connectives in contrast. *Languages in Contrast.* 12(2): 232-250.
- Meyer T., Grisot C. and Popescu-Belis A. (2013). Detecting Narrativity to Improve English to French Translation of Simple Past Verbs. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51th Annual Meeting of the Association for Computational Linguistics)*, Sofia, Bulgaria, pages 33-42.
- Meyer T., Popescu-Belis A., Hajlaoui N., Gesmundo A. (2012). Machine Translation of Labeled Discourse Connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego, CA.
- Zufferey S., Degand L., Popescu-Belis A. & Sanders T. (2012) - Empirical validations of multilingual annotation schemes for discourse relations. *Proceedings of ISA-8 (8th Workshop on Interoperable Semantic Annotation)*, Pisa, p.77-84.
- Meyer, T., Popescu-Belis, A. (2012). Using Sense-labeled Discourse Connectives for Statistical Machine Translation. *In Proceedings of the EACL 2012 Workshop on Hybrid Approaches to Machine Translation (HyTra)*, Avignon, France, pp. 129-138.
- Popescu-Belis A., Meyer T., Liyanapathirana J., Cartoni B. & Zufferey S. (2012). Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns. *Proceedings of LREC 2012*, May 23-25 2012, Istanbul.

References (continued)

- Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 909–919, Edinburgh.
- Guillou, L. 2013. Analysing lexical consistency in translation. In *Proceedings of the ACL Workshop on Discourse in Machine Translation*, Sofia, Bulgaria, pp. 10-18.
- Guillou L., Hardmeier C., Nakov P., Stymne S., Tiedemann J., Versley Y., Cettolo M., Webber B. & Popescu-Belis A. 2016. Findings of the 2016 WMT Shared Task on Cross-lingual Pronoun Prediction. *Proceedings of WMT 2016 (First Conference on Machine Translation)*, Berlin, Germany, p. 525–542.
- Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal.
- Jörg Tiedemann. 2010. Context adaptation in statistical machine translation using models with exponentially decaying cache. In Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, pages 8–15, Uppsala, Sweden.
- Ferhan Ture, Douglas W. Oard, and Philip Resnik. 2012. Encouraging consistent translation choices. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pages 417–426, Montréal, Canada.