MODERN: Modeling Discourse Entities and Relations for Coherent Machine Translation

Andrei POPESCU-BELIS[1], Jacqueline EVERS-VERMEUL[4], Mark FISHEL[3*], Cristina GRISOT[2*], Martin GROEN[4], Jet HOEK[4], Sharid LOAICIGA[2], Ngoc Quang LUONG[1], Laura MASCARELL[3], Thomas MEYER[1*], Lesly MICULICICH[1], Jacques MOESCHLER[2], Xiao PU[1], Annette RIOS[3], Ted SANDERS[4], Martin VOLK[3], Sandrine ZUFFEREY[4*]

** affiliation when working on the project*

[1] Idiap Research Institute, Martigny, Switzerland
[2] University of Geneva, Department of Linguistics
[3] University of Zürich, Institute of Computational Linguistics
[4] Utrecht Institute of Linguistics, The Netherlands

Motivation

- MT is efficient, has good coverage, is quite intelligible, but it always translates sentence by sentence and uses local features → it does not propagate information across sentences or even clauses
- Such information is crucial for the correct and coherent translation of complex sentences or entire texts
  - referring information: noun phrases, pronouns
  - verbs: tense, mode, aspect
  - discourse relations, as signaled by discourse connectives
  - other features, not targeted here: style, register, politeness
- This information is not (yet) accurately captured or used by mainstream MT systems (SMT or RBMT)

Example

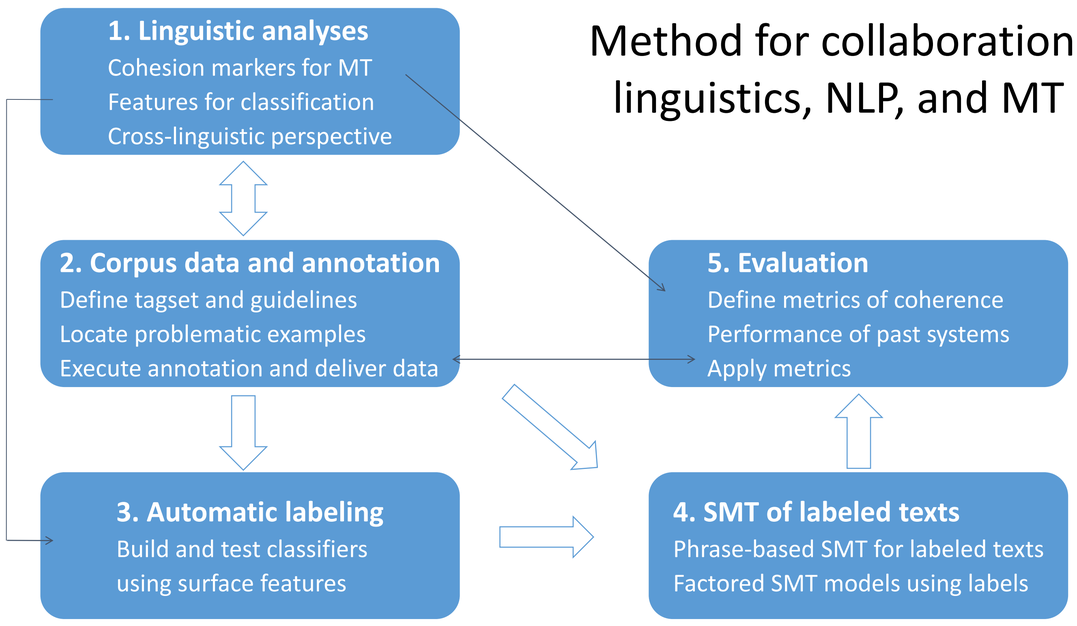
(two clauses)

	1. Connectives	2. Pronouns	3. Verb tenses		
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 for collaboration: linguistics, NLP, and MT



Support and related initiatives

- Swiss National Science Foundation, Sinergia program, grant n. 147653 www.idiap.ch/project/modern
- Continues COMTIS SNSF Sinergia project grant n. 127510 www.idiap.ch/project/comtis
- Resources available at www.idiap.ch/dataset
- Contributed to the creation of the **DiscoMT** workshops on **Discourse and Machine Translation** (in conjunction with ACL 2013 and EMNLP 2015)
- Involved in the **TextLink EU COST** network: **Structuring Discourse in Multilingual Europe**

Discourse connectives

- **Goal**: model the human translation of connectives, annotate parallel data, disambiguate source-side connectives automatically, use the result for SMT, and measure the improvement: EN → FR, DE, IT, AR
- **Cross-lingual studies of discourse connectives**
  Zufferey S. and Cartoni B. (2014). A multifactorial analysis of explicitation in translation. *Target*, vol. 26(3), p. 361-384.
  Zufferey S. & Cartoni B. (2012). English and French causal connectives in contrast. *Languages in Contrast*. 12(2): 232-250.
- **Annotation through translation spotting on Europarl**
  Cartoni B., Zufferey S., Meyer T. (2013) - Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *Dialogue and Discourse*, vol. 4, n. 2, pp. 65-86.
- **Automatic disambiguation and use for SMT with Moses**
  Meyer T., Hajlaoui N., and Popescu-Belis A. (2015). Disambiguating Discourse Connectives for Statistical Machine Translation. *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 23, n. 7, p.1184-1197.
- **Design of a reference-based evaluation metric (ACT)**
  Hajlaoui N. and Popescu-Belis A. (2013). Assessing the Accuracy of Discourse Connective Translations: Validation of an Automatic Metric. *Proc. of CICLing 2013*, Springer LNCS 7817, p.236-247.
- **Additional goal**: model the implicitation of discourse connectives, i.e. conveying a discourse relation without using an explicit discourse connective
- **Cross-lingual modeling**
  Hoek, J., and Zufferey, S. (2015). Factors influencing the implicitation of discourse relations across languages. *Proc. of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London.
  Hoek, J., Evers-Vermeul, J., & Sanders, T. (2015). The role of expectedness in the implicitation and explicitation of discourse relations. *Proc. of the Second DiscoMT workshop at EMNLP 2015*, Lisbon, Portugal.
- **Automatic selection of connectives to translate implicitly**
  Meyer T. and Webber B. (2013). Implicitation of Discourse Connectives in (Machine) Translation. In *Proc. of the First DiscoMT Workshop at ACL 2013*, Sofia, Bulgaria, pages 19-26.

Verb tenses

- **Goal**: model verb tense, mode and aspect; infer features for translation; study the EN → FR divergency of Simple Past translated e.g. as *imparfait, passé simple,* or *passé composé*
- **Cross-language modeling of verb TMA: linguistics and pragmatics**
  Grisot C., Moeschler J. and Cartoni B. (2012). Jusqu'où les temps verbaux sont-ils procéduraux? *Nouveaux cahiers de linguistique française*, vol. 30, p. 119-139.
  Grisot C. (2016). Temporal coherence in discourse: theory and application for Machine Translation. Blochowiak J. et al. eds., *Formal models in the study of language*, Springer.
- **Annotation of narrativity, disambiguation and use for MT**
  Grisot C. and Meyer T. (2014). Cross-Linguistic Annotation of Narrativity for English/French Verb Tense Disambiguation. *Proc. of LREC 2014 (9th International Conference on Language Resources and Evaluation)*, Reykjavik.
  Meyer T., Grisot C. and Popescu-Belis A. (2013). Detecting Narrativity to Improve English to French Translation of Simple Past Verbs. *Proc. of the First DiscoMT Workshop at ACL 2013*, Sofia, Bulgaria, pages 33-42.
- **Automatic alignment, disambiguation and MT of EN/FR verbs**
  Loaiciga S., Meyer T. and Popescu-Belis A. (2014). English-French Verb Phrase Alignment in Europarl for Tense Translation Modeling. *Proc. of LREC 2014 (9th International Conference on Language Resources and Evaluation)*, Reykjavik.

Pronouns

- **Goal**: improve pronoun translation by addressing translation divergencies, in particular, for EN → FR, the possible translations of *it* (as *il, elle, ça, ce,* etc.) and *they* (as *ils, elles,* etc.) – participated in the DiscoMT '15 and WMT '16 shared tasks
- **Modeling uncertainty on coreference resolution and definition of a coreference-aware decoder**
  Luong N.Q., Miculicich L., and Popescu-Belis A. (2015). Pronoun translation and prediction with or without coreference links. In *Proc. of the Second Workshop on Discourse in Machine Translation*, p. 94–100, Lisbon, Portugal.
  Luong N.Q. and Popescu-Belis A. (in preparation). Improving Pronoun Translation by Modeling Coreference Uncertainty.
- **Defining a pronoun language model to re-rank SMT output**
  Luong N. Q. and Popescu-Belis A. (2016). A Contextual Language Model to Improve Machine Translation of Pronouns by Re-ranking Translation Hypotheses. *Proc. of EAMT 2016 (19th Annual Conference of the European Association for Machine Translation)*, Riga, Latvia, 13 p.
- **Validating a reference-based evaluation metric**
  Miculicich L. and Popescu-Belis A. (in preparation). Validation of an Automatic Metric for the Accuracy of Pronoun Translation (APT).

Noun phrases

- **Goal**: improve the consistency of noun phrase translation, in particular for repeated nouns, including compounds, for DE → EN/FR & ZH → EN
- **Enforcing consistent translation of compound elements: *XY … Y***
  Mascarell L., Fishel M., Korchagina N., and Volk M. (2014). Enforcing Consistent Translation of German Compound Coreferences. *Proc. of KONVENS 2014 (12th German Conference on Natural Language Processing)*, Hildesheim, Germany.
  Pu X., Mascarell L., Popescu-Belis A., Fishel M., Luong N.Q., and Volk M. (2015). Leveraging Compounds to Improve Noun Phrase Translation from Chinese and German. *Proc. of the ACL-IJCNLP 2015 Student Research Workshop*, Beijing, China.
- **Enforcing consistent translation of nouns: *X … X → N_1 … N_2***
  Pu X. and Mascarell L. (in preparation). Consistent translation of repeated nouns using syntactic and semantic cues.