Using Discourse Structure as Textual Context for Statistical Machine Translation: the COMTIS Project

Andrei Popescu-Belis
Idiap Research Institute

September 11, 2012

META-NET Workshop: Machine Translation and Multimodal Contexts
Special session at ICANN 2012, Lausanne, Switzerland, Sept. 11-14, 2012

COMTIS: SNF Sinergia project

- "Improving the **co**herence of **m**achine **t**ranslation output by modeling **i**ntersentential relation**s**"

- Idiap + two groups at the University of Geneva
  - with: B.Cartoni, C.Grisot, N.Hajlaoui, J.Henderson, J.Liyanapathirana, P.Merlo, T. Meyer, J.Moeschler, S.Zufferey
- Three-year project: March 2010 – February 2013
  - likely extended (Aug. 2013) and continued (Aug. 2014)

Motivation

			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

COMTIS considers the context of longer-range dependencies: translation in green

Main idea of the project

- MT can be improved by giving it information about context across sentences
  - text-level information or, rather, *discourse-level*
  - might involve intra-sentence dependencies as well

- How to use context
  - extract contextual features from source text
  - design MT systems that can use these features

How to proceed?

1. Define the contextual phenomena to target
    - linguistic analysis, relevance to MT, tractability
2. Create data for training and evaluation
    - also for corpus linguistics
3. Build classifiers for each phenomenon
    - not necessarily perfect, but useful for MT
4. Adapt MT systems to use classifiers' output
5. Evaluate the improvement of MT

What to model and annotate

languages: EN, FR, DE, IT

What makes discourse discourse?

- "Text-level" exists because texts are generally coherent
  - coherence is ensured by cohesion markers
- What cohesion markers might help SMT in COMTIS?
  - tense: modeling (PhD at UniGe)
  - connectives: modeling + annotation + classifiers (PhD at Idiap)
  - pronouns: annotation + post-editing (intern/PhD at Idiap)
  - lexical choice: *later*
  - style / register: *later*
- Cohesion markers are long-range or intersentential
  - current SMT systems translate sentence-by-sentence
  - some commercial rule-based MT consider text-level domains

Examples (1)

- Discourse connectives
  - SOURCE: Why has no air quality test been done on this particular building *since* we were elected? (Europarl)

    Ref: Comment se fait-il qu'aucun test de qualité de l'air n'ait été réalisé dans ce bâtiment *depuis* notre élection?

    SMT: Pourquoi aucun test de qualité de l' air a été réalisé dans ce bâtiment *car* nous avons été élus ?

  - SOURCE: While no-one wants to see public demonstration, I have to say I understand the anxiety and share their concern. (Europarl)

    Ref: *Alors que* personne ne veut voir de manifestations publiques, je dois dire que je comprends leur anxiété et que je partage leur inquiétude.

    SMT: *Bien que* personne ne veut voir la démonstration publique, je dois dire que je comprends l'inquiétude et de partager leurs préoccupations.

Examples (2)

- Tense
  - SOURCE: Grandmother drank three cups of coffee a day.

    Ref: Grand-maman *buvait* trois tasses de café par jour.

    SMT: Grand-mère *a bu* trois tasses de café par jour.
  - SOURCE: Je me lève à cinq heures depuis 20 ans.

    Ref: I *have been waking up* at five o'clock for the last 20 years.
    SMT: I *get up* at five in the last 20 years.

- Pronouns
  - 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 …

Some achievements from the first two years

Modeling verb tense

- How to label verb tenses to ensure that they are coherently translated?
  - depends on the language pair
  - must be tractable for NLP
    - existing linguistic theories of tense are complex
  - what features are useful to compute labels?

- A model for the translation of EN simple past into FR (mainly *passé simple* vs. *imparfait*) has been proposed and justified (theoretically and empirically)
  - pilot annotation of resources, more needed for training

Translating EN simple past

- Proposed label: ‘narrative’ vs. ‘non-narrative’
  - must be assigned globally, at the text level
- Proposed impact of label on MT
  - simple past ‘narrative’ → *passé simple* (or *composé*)
  - simple past ‘non-narr.’ → *imparfait*
- How to assign this label automatically?
  - *we don’t know yet, but will look at training data*
- This is really a simplified view
  - more labels, e.g. ‘subjective’ or not
  - more EN and FR tenses

Modeling and annotating discourse connectives

- Existing theories and annotated resources (mainly EN)
  - PDTB: complex hierarchy of possible senses of connectives
    - difficult to annotate, not necessarily relevant to MT

- In COMTIS, annotation through **translation spotting**
  - annotators only identify the human translation of each connective in a parallel corpus (Europarl)
  - for each connective type, observed translations are clustered into *a posteriori* "senses" relevant to MT
    - compact set of labels, cheaper to annotate
    - done for English/French, English/German/Italian, Arabic in progress
- Example
  - PDTB: *while* has 21 possible composite labels
  - COMTIS: *while* signals either a contrast, a concession, or has a temporal meaning (durative, temporal, or causal)

Annotations of connectives in COMTIS

Lexical items	A priori senses	A posteriori senses	N.S.	F.S.
EN CONNECTIVE			*Total: 1,246*	
although	contrast, concession	contrast, concession	197	183
even though	contrast, concession	contrast, concession	212	191
since	temporal, causal	temporal, causal_known_relation, causal_new_relation, causal_other	423	423
though	contrast, concession	contrast, concession	200	155
while	contrast, concession, comparison, temporal	contrast, concession, contrast_and_temporal, temporal_durative, temporal_punctual, temporal_causal	499	294
FR CONNECTIVE			*Total: 817*	
dans la mesure où	causal, explanation	causal_and_explanation	175	150
pourtant	contrast_and_concession	contrast_and_concession	312	250
alors que	contrast, temporal	contrast, temporal, temporal_and_contrast	423	366
bien que	concession	concession_and_contrast	55	51

Automatic labeling of connectives

- Classification problem: for each discourse connective
  - given features extracted from the text
  - determine its most probable label (“sense”)
  - using MaxEnt, decision trees, etc.
    - trained on manually labeled data (PDTB or COMTIS)
    - tested on unseen data or plugged into an SMT system

- Features
  - standard
    - token, capitalization, POS tag, parent syntactic class, punctuation
    - first/last word/POS of previous/current clause
  - novel
    - similarity/antonymy for word pairs in the two clauses (WordNet)
    - features related to temporal relations (Tarsqi Toolkit)
    - candidate translation from a baseline SMT system

Example of results on PDTB

Connective	Number of occurrences and senses				F1 Score PT	F1 Score PT+
	Training set: total and per sense		Test set: total and per sense			
after	507	456 As, 51 As/Ca	25	22 As, 3 As/Ca	0.66	1.00
although	267	135 Cs, 118 Ct, 14 Cp	16	9 Ct, 7 Cs	0.60	0.66
however	176	121 Ct, 32 Cs, 23 Cp	14	13 Ct, 1 Cs	0.33	1.00
indeed	69	37 Cd, 24 R, 3 Ca, 3 E, 2 I	*2	2 R	*0.50	*0.50
meanwhile	117	66 Cj/S, 16 Cd, 16 S, 14 Ct/S, 5 Ct	10	5 S, 5 Ct/S	0.32	0.53
nevertheless	26	15 Ct, 11 Cs	6	4 Cs, 2 Ct	0.44	0.66
nonetheless	12	7 Cs, 3 Ct, 2 Cp	*1	1 Cs	*1.00	*1.00
rather	10	6 R, 2 Al, 1 Ca, 1 Ct	*1	1 Al	*0.00	*0.00
since	166	75 As, 83 Ca, 8 As/Ca	9	4 As, 3 Ca, 2 As/Ca	0.78	0.78
still	114	56 Cs, 51 Ct, 7 Cp	13	9 Ct, 4 Cs	0.60	0.66
then	145	136 As, 6 Cd, 3 As/Ca	6	5 As, 1 Cd	0.83	1.00
while	631	317 Ct, 140 S, 79 Cs, 41 Ct/S, 36 Cd, 18 Cp	37	19 Ct, 10 S, 4 Cs, 4 Ct/S	0.93	0.96
yet	80	46 Ct, 25 Cs, 9 Cp	*2	2 Ct	*0.5	*1.00
Total	**2,320**	–	**142**	–	**0.57**	**0.75**

Al: alternative, As: asynchronous, Ca: cause, Cd: condition, Cj: conjunction, Cp: comparison, Cs: concession, Ct: contrast, E: expansion, I: instantiation, R: restatement, S: synchrony

Integration with MT

- How to train SMT to use labeled connectives?
- Several methods have been studied
  - replace in the system's phrase table all unambiguous occurrences of the connective with the labeled connective
  - train the system on manually or on automatically labeled data (e.g., *while* becomes *while_Temporal*)
  - combine contextual features into factored MT models
  - train over multiplied data in proportion to the label prob.
- Also: use a modified SMT system only when the connective labeler is confident enough

Sample results

- Modified phrase table
  - tested on ~10,000 instances of connectives (5 types)
  - **34%** improved, **20%** degraded, **46%** unchanged [SAMPLE]
- Trained on manually labeled data
  - **26%** improved, **8%** degraded, **66%** unchanged [SAMPLE]
- Trained on automatically labeled data
  - **18%** improved, **14%** degraded, **68%** unchanged [SAMPLE]
  - smaller improvement, but cheaper and larger data
- Thresholding based on labeler's confidence
  - experimented with two connectives until now
  - improvement of **0.2**-**0.4** BLEU points (quite significant)

What about global output quality?

- It depends on how we measure it
  - traditional BLEU metric: n-gram automatic comparison of a candidate text with one or more reference translations
- COMTIS
  - contextual factors are not frequently determinant
  - so impact on BLEU should be small (and it is)
  - goal is at least not to decrease BLEU scores
- Need for specific automatic metrics
  - still reference-based, but sensitive to sparse phenomena

Towards new evaluation metrics

- Goal: automatic procedure to count how many connectives were 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 translation(s)
    - how *C* is translated in the candidate translation
  - compare the two translations of *C*
    - identical / "synonymous" / incompatible / missing

- ACT empirically tested: within 1-5% of human ratings
  - can also be used to spot litigious sentences, which are given to human assessment (10-20% of all sentences)

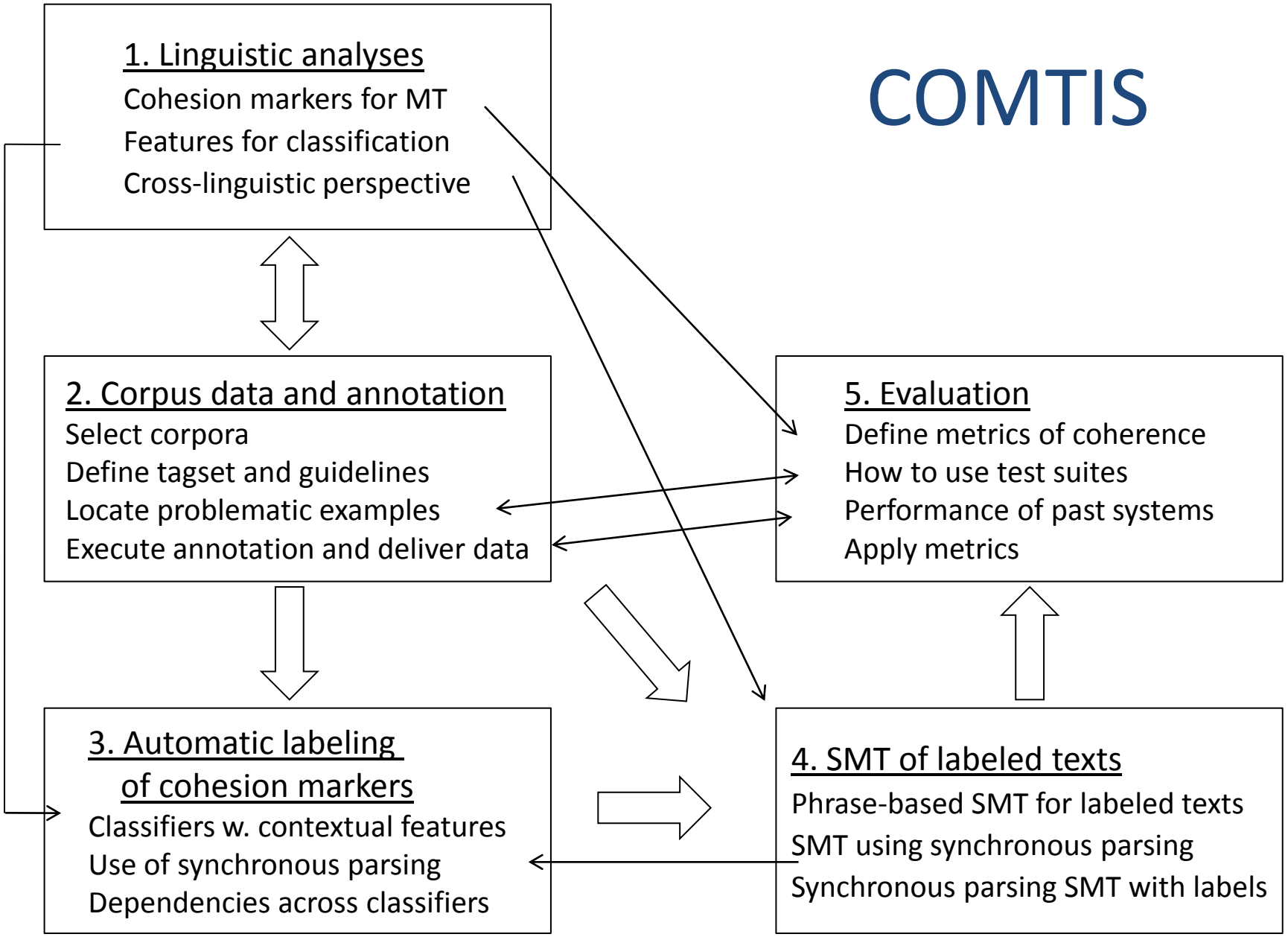
Recent results, on WMT10 data

- Factored models (with Moses SMT)
  - source factors: POS tags, labeled connectives (DL), or both (POS+DL)
  - phrase-based or hierarchical SMT models
- Non-factored models: multiplied data based on labels' probabilities

Translation model	**SMT system**	**BLEU**	ACT_a	ACT_{a5}	ACT_m
Factored phrase-based	POS + DL	22.19	70.7	86.1	82.1
	DL	21.69	70.0	85.2	80.7
	POS	22.26	67.9	81.2	76.4
	Baseline	21.71	65.0	77.8	73.6
Factored hierarchical	DL	19.20	67.9	78.5	77.1
	Baseline	19.31	63.6	74.8	74.3
Phrase-based with label probabilities	LPD	21.60	69.4	82.0	78.5
	Baseline	21.30	68.8	81.1	79.2

Wrap up

COMTIS


1. Linguistic analyses
Cohesion markers for MT
Features for classification
Cross-linguistic perspective
2. Corpus data and annotation
Select corpora
Define tagset and guidelines
Locate problematic examples
Execute annotation and deliver data
5. Evaluation
Define metrics of coherence
How to use test suites
Performance of past systems
Apply metrics
3. Automatic labeling of cohesion markers
Classifiers w. contextual features
Use of synchronous parsing
Dependencies across classifiers
4. SMT of labeled texts
Phrase-based SMT for labeled texts
SMT using synchronous parsing
Synchronous parsing SMT with labels

Perspectives

- Make progress on all tasks
  - more resources, better integration with MT, process new phenomena, improve evaluation
- Towards a proof-of-concept
  - text-level processing is efficient enough for MT
  - it can be efficiently combined with MT
- Check www.idiap.ch/comtis for more details

Some COMTIS publications

Meyer T., Popescu-Belis A., Hajlaoui N. & Gesmundo A. (in press) - Machine Translation of Labeled Discourse Connectives. *Proceedings of AMTA 2012 (10th Conference of the Association for Machine Translation in the Americas)*, San Diego, CA, 10 p.

Meyer T. & Popescu-Belis A. (2012) - Using Sense-labeled Discourse Connectives for Statistical Machine Translation. *Proceedings of the EACL 2012 ESIRMT-HyTra Workshop (Hybrid Approaches to MT)*, Avignon, p.129-138.

Cartoni B., Zufferey S., Meyer T. & Popescu-Belis A. (2011) – How Comparable are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives. *Proceedings of BUCC 2011 (4th Workshop on Building and Using Comparable Corpora, at ACL-HLT 2011)*, Portland, OR, p.78-86.

Meyer T., Popescu-Belis A., Zufferey S. & Cartoni B. (2011) - Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation. *Proceedings of SIGDIAL 2011 (12th annual SIGdial Meeting on Discourse and Dialogue)*, Portland, OR, p.194-203.