Manual and Automatic Labeling of Discourse Connectives for Machine Translation

Andrei Popescu-Belis

Idiap Research Institute, Martigny, Switzerland

TextLink Second Action Conference

Budapest, 11 April 2016

A limitation of machine translation

- MT is efficient, has good coverage, is quite intelligible, *but* it always translates sentence by sentence, using local features
  - it does not propagate information across sentences or clauses
- Still, such information is crucial for the correct and coherent translation of complex sentences or entire *texts*
  - referring information: noun phrases (terms), pronouns
  - verbs: tense, mode, aspect
  - global features: style, register, politeness
  - **discourse relations, as signaled by discourse connectives**
- This information is not (yet) accurately captured or used by mainstream MT systems, statistical or rule-based

Desired improvements

			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

How to achieve these improvements?

1. Define and analyze the phenomena to target
   - design theoretical models accessible to automatic processing
2. Create data for system development & evaluation
   - labeling instructions + annotation of data sets
   - validate linguistic models through corpus studies
3. Perform automatic recognition/disambiguation
   - automatic classifiers, e.g. based on machine learning from annotated data, using surface features
4. Modify MT systems to use automatic labels
5. Measure changes in connective translation

Joint effort between five teams

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

- Funded by the Swiss National Science Foundation since 2010

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

  www.idiap.ch/project/comtis   www.idiap.ch/project/modern

  MODERN: Modeling discourse entities and relations for coherent MT

- People collaborating in these projects
  - **Idiap Research Institute, NLP group**: APB, Thomas Meyer, Quang Luong, Najeh Hajlaoui, Xiao Pu, Lesly Miculicich, 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, Annette Rios, Laura Mascarell
  - **Utrecht Institute of Linguistics**: Ted Sanders, Jacqueline Evers-Vermeul, Martin Groen, Jet Hoek

Plan of the talk

1. Motivation
2. Definition of labels for discourse connectives
3. Annotation of discourse connectives
4. Automatic disambiguation
5. Integration with statistical MT
6. Conclusion and perspectives

Note

- translation from English into French (and German)
- genres: parliamentary debates (Europarl), news (Wall Street Journal/PTDB)

1. MOTIVATION

Issues with discourse connectives in MT

- 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.

Importance of discourse connectives to machine translation (1/2)

- "Small words, big effects"
  - signal discourse relations between sentences or clauses
    - addition, temporal, cause, condition, contrast, etc.
- Assumptions made in our studies
  - discourse relations are preserved in translation
  - implicitation (e.g., *since* → Ø) and explicitation (e.g., Ø → *en effet*) of discourse connectives are not considered

Importance of discourse connectives to machine translation (2/2)

- Challenge to translation: connectives may signal different relations, which may be translated differently
  - *since* causal or temporal: French *puisque* or *depuis que*
  - *while* concessive or contrastive or temporal: French *bien que* or *mais* or *pendant que*

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

  ➔ low coherence or readability

2. DEFINITION OF LABELS FOR DISCOURSE CONNECTIVES

Modeling and annotating discourse connectives

- Main existing theories
  - Rhetorical Structure Theory (Mann and Thompson)
  - Discourse Representation Theory (Asher et al.)
  - Cognitive approach to Coherence Relations (Sanders et al.)

- Annotation-oriented approach: Penn Discourse Treebank (PDTB) (Prasad, Webber, Joshi et al.)

- PDTB: complex hierarchy of possible senses of connectives
  - specified for English, then used e.g. for Arabic, Hindi, Italian (with some adaptations)
  - PDTB-style taxonomies defined for Chinese, Czech, French

Requirements for labels to be usable with MT

- Availability of parallel corpora with labeled discourse connectives on the source-side

- PDTB: English, 1 M tokens, 18,459 explicit connectives
  - not parallel: no available translations
  - rather complex hierarchy of senses of connectives
    - not all distinctions are relevant to MT (EN/FR)
    - costly to annotate

- Two possible solutions
  1. Translate PDTB (WSJ) texts into French (10¢/word)
  2. Annotate new parallel data, such as Europarl

Some annotation attempts

- Classical manual annotation of the senses: trained annotators were asked to label connectives in context with appropriate senses
- Two experiments showed low inter-coder agreement, as well as significant effort and time required

while

- opposition / concession / comparison / temporal → κ = 0.56

alors que

- background / contrast → κ = 0.43

→ Need for a quicker method and a simpler tag set

idiap RESEARCH INSTITUTE

3. ANNOTATION OF DISCOURSE CONNECTIVES

1. “Transpotting” of discourse connectives

Translation spotting: find the translations

While we have a duty to tackle this problem within EU waters, ultimately this is a problem which requires international action.	Bien que nous ayons le devoir de traiter ce problème au niveau des eaux de l'UE, il s'agit en dernier ressort d'un problème qui exige des actions au niveau international.
No wonder Richard Holbrooke recently boasted that Europe slept **while** President Clinton resolved a particular European crisis.	Il n'y a dès lors rien d'étonnant à ce que M. Richard Holbrooke nous ait récemment nargué en disant que l'Europe dormait pendant que le président Clinton résolvait une crise européenne particulière.
…	…

Performed on parallel sentences from Europarl

2. Clustering of the annotated translations to define new, application-oriented labels

Translations of *while*	**Nb.**	**%**
alors que	91	18.24%
[gerund]	85	17.03%
[paraphrase]	72	14.43%
si	54	10.82%
[no translation]	41	8.22%
tandis que	39	7.82%
même si	33	6.61%
bien que	26	5.21%
s'il est vrai que	14	2.81%
tant que	10	2.00%
pendant que	5	1.00%
puisque	5	1.00%
lorsque	4	0.80%
mais	4	0.80%
…	…	…
Total	**499**	**100%**

Labels of clusters	
Contrast/Temporal	C
Concession/Condition	A
Contrast	B
Concession	A
Concession	A
Concession/Condition	A
Temporal/Condition	D
Temporal/Duration	E
Temporal/Punctual	F
Contrast	B
Note: PDTB has 21 labels, vs.	**6.**

3. Projection of the cluster label onto the source discourse connectives

While we have a duty to tackle this problem within EU waters, ultimately this is a problem which requires international action.	**Bien que** nous ayons le devoir de traiter ce problème au niveau des eaux de l'UE, il s'agit en dernier ressort d'un problème qui exige des actions au niveau international.	bien que	concession
No wonder Richard Holbrooke recently boasted that Europe slept **while** President Clinton resolved a particular European crisis.	Il n'y a dès lors rien d'étonnant à ce que M. Richard Holbrooke nous ait récemment nargué en disant que l'Europe dormait **pendant que** le président Clinton résolvait une crise européenne particulière.	pendant que	temporal/ duration
…	….	…..	

Advantages and drawbacks of translation spotting

- Advantages
  - simplicity of the scheme: quicker and more reliable manual annotation / potentially easier automatic one
  - empirically grounded
  - adapted to the translation problem
    - the labels are those that make a difference in translation

- Drawbacks
  - <u>different</u> senses rendered by the <u>same</u> connective in translation are not distinguished
  - specificity to a given language pair
    - if we transpot the same EN source using either EN/FR or EN/DE alignments, the labels may differ (actually not much)

Annotated connectives and senses

	English connectives	**2379**
as	CAUSAL, CONCESSION, COMPARISON, TEMPORAL (ALSO: PREPOSITION)	599
although	CONTRAST, CONCESSION	183
even though	CONTRAST, CONCESSION	191
meanwhile	CONTRAST, TEMPORAL	131
since	TEMPORAL, TEMPORAL_AND_CAUSAL, CAUSAL_KNOWN_RELATION, CAUSAL_NEW_RELATION, CAUSAL_OTHER	558
though	CONTRAST, CONCESSION	155
while	CONTRAST, CONCESSION, CONTRAST_AND_TEMPORAL, TEMPORAL_DURATIVE, TEMPORAL_PUNCTUAL, TEMPORAL_CONDITIONAL	294
yet	ADVERB, CONTRAST, CONCESSION	403
	French connectives	**817**
alors que	CONTRAST, TEMPORAL, TEMPORAL_AND_CONTRAST	366
bien que	CONTRAST, CONCESSION	51
dans la mesure où	CONDITION, EXPLANATION	150
pourtant	CONTRAST, CONCESSION	250

4. AUTOMATIC DISAMBIGUATION (OR LABELING)

Automatic labeling of connectives

- Classification problem
  - for each discourse connective
    - automatically extract features from the text
    - use an automatic classifier to determine its label (sense)
- Classifiers can be
  - designed *a priori,* e.g. by writing a set of rules
  - learned (trained, optimized) from labeled data

Training and test sets from Europarl (with translation spotting) and PDTB

Connective	Training set EP	Training set PDTB	Training set Distribution of labels (%)
although	168	312	Ct: 68.9; Cs: 31.1
however	348	450	Ct: 47.8; Cs: 52.2
meanwhile	102	177	Ct: 77.3; T: 22.7
since	339	174	Ca: 38.7; T: 59.6; T/Ca: 1.7
(even) though	276	306	Ct: 33.3; Cs: 66.7
while	236	744	Ct: 14; Cs: 23; T: 15; T/Ct: 46.6; T/Cd: 1.4
yet	326	99	Ct: 23.2; Cs: 29.8; Adv: 47
Total	1795	2262	–

Testing set EP	Testing set PDTB	Testing set Distribution of labels (%)
15	16	Ct: 48.4; Cs: 51.6
70	35	Ct: 47.6; Cs: 52.4
28	14	Ct: 76.2; T: 23.8
82	10	Ca: 30.4; T: 67.4; T/Ca: 2.2
69	14	Ct: 33.7; Cs: 66.3
58	37	Ct: 22.8; Cs: 33.7; T: 9.8; T/Ct: 30.4; T/Cd: 3.3
77	2	Ct: 30.4; Cs: 19; Adv: 50.6
399	128	–

T: temporal
Ct: contrast
Cs: concession
Cd: conditional
Ca: causal
Adv: adverb

Features for the automatic disambiguation of connectives

- Extracted from the current and the previous sentences

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 (token, with capitalization information), punctuation, context words (first/last word and POS), context tree structures (parent syntactic class), auxiliary verbs
- WordNet antonymy features
  - similarity scores (WordNet distance) and antonyms of word pairs from the clauses
- TimeML features
  - temporal relations extracted with the Tarsqi toolkit by Verhagen and Pustejovsky (2008)
- discourse relation features
  - discourse relations from RST-style discourse parser by Soricut and Marcu (2003)
- polarity features
  - using a polarity lexicon, count positive and negative words, account for negation
- translational features
  - candidate translation from baseline MT (e.g. *tandis que*), "sense", position

Experiments

- Input data: extracted features + labels
  - subsets of Europarl (transpot) and PTDB (with conversion of labels)
- Supervised learning: trained a classifier on the input data
  - NB: training = find a classifier which would, using only the features, output labels as similar as possible to those annotated by people
  - considered several possible classifiers from the WEKA toolkit
    - Maximum Entropy (logistic regression), Decision Trees, Bayesian, etc.
- Test data with manual labels, or cross-validation
  - c.v. = permute training/test sets *N* times, average scores on test sets

Performance of automatic connective labeling

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 (F1-score: average of recall and precision per class)
  - scores generally compare well to inter-annotator agreement levels (80-90%) and to the state of the art
  - using *all features* is the best option

5. INTEGRATION WITH MACHINE TRANSLATION

How do we use labeled connectives for MT?

- State of the art machine translation systems
  - direct rule-based, e.g. Systran: costly to build, hard to modify
  - statistical: phrase-based or hierarchical, e.g. Moses toolkit
    - easy to build from parallel data, though with high computational costs
    - easy to modify, e.g. by adding other "factors" than TM and LM

- How do we constrain the translation produced by SMT?
  - brute force post-editing

    → not enough specific, leads to many mistakes
  - combination with statistical MT

    → let SMT learn and then use the translations of labeled connectives along with its own translation model and language model

How do we measure the changes in connective translation?

- Measuring translation quality
  - subjective (human) measures: fluency, fidelity → expensive
  - 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 on BLEU scores

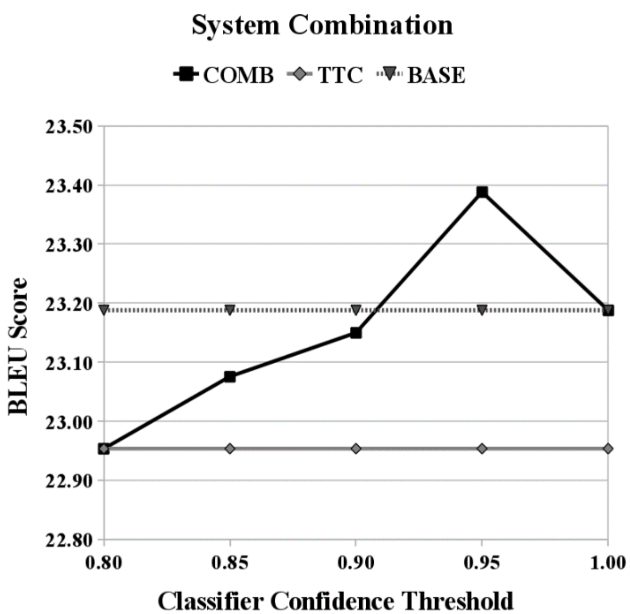
- Count how many connectives are correctly translated: ACT metric [Accuracy of Connective Translation]
  - given a source sentence with a discourse connective *C*
  - use automatic alignment to find out:
    - how *C* is translated in the reference and in the candidate translations
  - count the translations: (1) identical (2) "synonymous" (3) incompatible (4, 5, 6) absent (on each side)

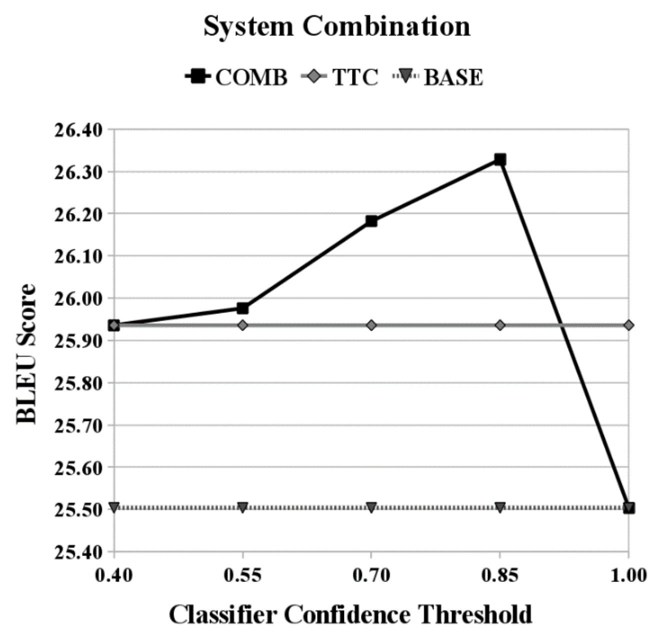
Learning an SMT system from data with labeled discourse connectives

- First method: “concatenated labels”
  - append to each occurrence of a discourse connective its label
    - e.g. *while* → *while_Temporal*
  - this creates new “words”: their translations can be learned
- Training data (parallel): two options
  1. Manually-labeled data: reliable but low volume available
  2. Automatically-labeled data: abundant but imperfect
- Results for each option
  1. 26% improved, 8% degraded, 66% unchanged
  2. 18% improved, 14% degraded, 68% unchanged

Exploiting the confidence of labels

- Thresholding based on automatic labeler's confidence
  - use the connective-specific SMT system (concatenated words, trained on automatically-labeled data) when the connective labeler is confident enough, otherwise use the baseline system
- Results (left: *although*, right: *since*)
  - improvement of 0.2-0.4 BLEU points: small but significant





Labels on discourse connectives used as “factors” in SMT

- Second method: use Factored Models as implemented in Moses
  - word-level linguistic labels function as separate translation features
  - a model of labels is learned when training, then used when decoding
  - the labels are still assigned automatically on a large data set

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

6. CONCLUSIONS & PERSPECTIVES

Main findings

- Manual annotation of discourse connectives
  - translation-oriented set of labels
  - translation spotting as a cost-effective annotation method
  - made available annotation of 2,379 EN connectives and 817 FR ones
- Automatic labeling of connectives
  - new features including inter-sentential, semantic ones
  - reached or improved state-of-the-art labeling performance
- Translation of connectives by using automatic labeling in SMT
  - NB: strict evaluation metric: identity to a human translation
  - improved the fully-automatic end-to-end translation

    ➔ training SMT on manual annotations better than on automatic ones

    ➔ when no source-side manual annotations are available, training SMT on automatic annotations still brings improvements

Challenges for the future: discourse connectives

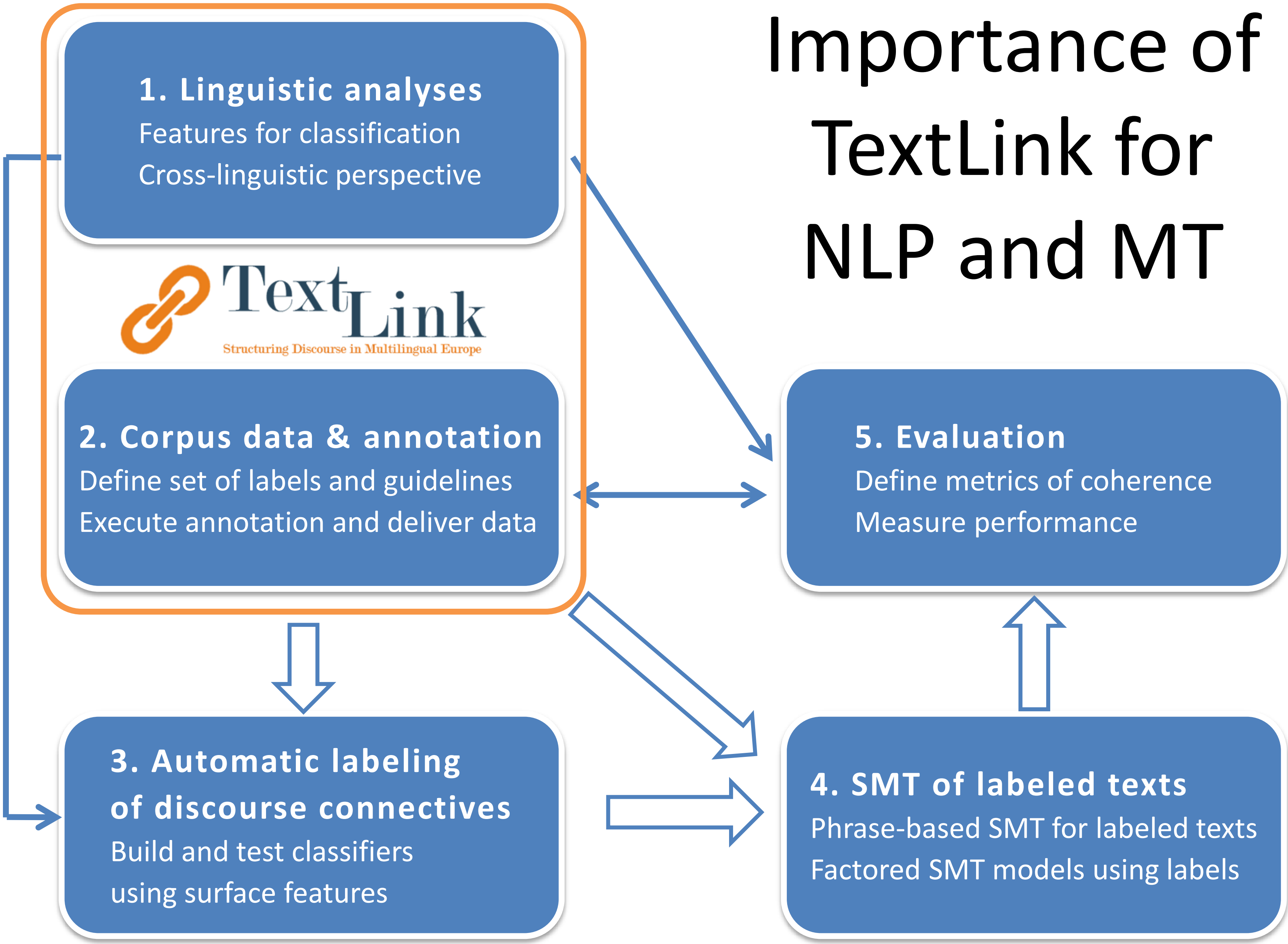
- Improve machine translation of (explicit) connectives
  - larger amounts of training data
    - from various sources, e.g. using mappings across sets of labels
  - more expressive and better grounded labels
  - more informative features for automatic classification

- Automatic implicitation / explicitation of connectives
  - better understanding of the factors governing them
  - implicitation
    - decide what source-side connectives not to translate
  - explicitation
    - find the discourse relation or *implicit connective* on the source side
    - decide *how* and *where* to express it on the target side

Challenges for the future: discourse-level machine translation

- Apply the method to other cohesion marks
  - verb tenses: already attempted on EN/FR Simple Past
  - consistency of repeated nouns, including compounds
  - pronoun divergencies (*it* → *il* / *elle* / *c'* / *ce* / *cela* / …)
  - what are other promising phenomena?
- New methods to use discourse information for MT
  - how can we efficiently integrate several complex and heterogeneous knowledge sources into SMT?

Importance of TextLink for NLP and MT



References

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), p.1184-1197.

Meyer T. (2015) - *Discourse-level features for statistical machine translation*, PhD thesis, École polytechnique fédérale de Lausanne (EPFL), n. 6501, 2015.

Cartoni B., Zufferey S., Meyer T. (2013) - "Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique". *Dialogue & Discourse* : Beyond semantics: the challenges of annotating pragmatic and discourse phenomena. Vol. 4, No. 2, pp. 65-86.

Zufferey S. & Cartoni B. (2012) - English and French causal connectives in contrast. *Languages in Contrast.* 12(2): 232-250.

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.

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, Turkey.