Document-level Automatic MT Evaluation based on Discourse Representations

Elisabet Comelles and **Irene Castellón** (UB),
Jesús Giménez and **Lluís Màrquez** (UPC), and **Victoria Arranz** (ELDA)
UB GRIAL Research Group (Universitat de Barcelona) and ELDA/ELRA
UPC TALP Research Center (Universitat Politècnica de Catalunya)

Abstract

This paper describes the joint participation of Universitat Politècnica de Catalunya and Universitat de Barcelona at the Metrics MaTr 2010 evaluation challenge, in collaboration with ELDA/ELRA. Our work is aimed at widening the scope of current automatic evaluation measures from sentence to document level. Preliminary experiments, based on an extension of the metrics by Giménez and Màrquez (2009) operating over discourse representations, are presented.

Motivation

- **Problem:** Current automatic evaluation methods for Machine Translation (MT) operate all, without exception, at the segment level → *ignoring cross-sentential/discursive phenomena*.
- **Our Proposal:** We suggest widening the scope of evaluation methods by means of *genuine document-level measures based on Discourse Representations*.
  - We take advantage of the availability of:
    - Automatic linguistic processors which provide detailed discourse-level representations of text, e.g., the C&C Tools [CCB07]
    - Document-structured test beds from recent MT evaluation campaigns

Why Discourse Representations?

- They provide a *semantic representation*: events, entities, roles (agent, patient, theme, location, time, etc), relations (disjunction, implication, negation, question, propositional attitude, etc.).
- They capture *cross-sentence information* (e.g., anaphoric relations, discourse markers) and allow us to trace links across sentences between the different facts and entities appearing in them, e.g.:
  - the connection between a possessive adjective and a proper noun or a subject pronoun (e.g., *"Maria bought a new sweater. Her new sweater is blue."*).
  - the link between a demonstrative pronoun and its referent (e.g., *"He developed a new theory on grammar. However, this is not the only theory he developed"*).
  - the relation between a main verb and an auxiliary verb in certain contexts (e.g., *"Would you like more sugar? Yes, I would"*)
  - discourse markers:
    - "Moreover", "Furthermore", "In addition" indicate that the upcoming sentence adds more information to the discourse.
    - "However", "Nonetheless", "Nevertheless" show contrast with previous ideas.
    - "Therefore", "As a result", "Consequently" show a cause and effect relation.
    - "For instance", "For example" clarify or illustrate the previous idea.

Example 1 (a). *"Maria bought a new sweater. Her new sweater is blue."*

$$\left(\begin{array}{|l} X_0\ X_1 \\ \hline \text{a1\_female}(X_0) \\ \text{named}(X_0\text{,maria,per}) \\ \text{n1\_sweater}(X_1) \\ \text{r1\_of}(X_1\text{,}X_0) \\ \text{a1\_new}(X_1) \end{array}\right) \wedge \left(\begin{array}{|l} X_2\ X_3 \\ \hline \text{n1\_sweater}(X_2) \\ \text{n1\_event}(X_3) \\ \text{a1\_new}(X_2) \\ \text{v1\_buy}(X_3) \\ \text{r1\_agent}(X_3\text{,}X_0) \\ \text{r1\_patient}(X_3\text{,}X_2) \end{array}\right) \wedge \left(\begin{array}{|l} X_4 \\ \hline \text{n1\_event}(X_4) \\ \text{v1\_blue}(X_4) \\ \text{r1\_theme}(X_4\text{,}X_1) \end{array}\right))$$

- A female person named "Maria" who owns a new sweater exists.
- This person ("Maria") bought a new sweater. (fails to capture that it is the same sweater!)
- The new sweater owned by "Maria" is blue.

Metric Description

- As a first proposal, instead of elaborating on novel similarity measures, we have borrowed and extended the Discourse Representation *(DR)* metrics defined by Giménez and Màrquez (2009):

  DR-STM — *Semantic Tree Matching*. This metric is similar to the Syntactic Tree Matching metric defined by Liu and Gildea (2005), in this case applied to DRSs instead of constituent trees. All semantic subpaths in the candidate and reference trees are retrieved. The fraction of matching subpaths up to a given length (length is 4 in our experiments) is computed.

  DR-$O_r(\star)$ — *Average lexical overlap* between discourse representation structures of the same type. Overlap is measured according to the formulae and definitions by Giménez and Màrquez (2007).

  DR-$O_{rp}(\star)$ — Average morphosyntactic overlap, i.e., between grammatical categories –parts-of-speech– associated to lexical items, between discourse representation structures of the same type.
- Discourse representations are obtained using the Boxer component [Bos08] from the C&C Tools [CCB07].
- For porting these metrics to the document-level (DR_{doc}), we simply run the C&C Tools in a document-by-document fashion, instead of sentence by sentence.

Example 1 (b). *"Maria bought a new blue sweater."*

$$\left(\begin{array}{|l} X_0\ X_1 \\ \hline \text{n1\_time}(X_0) \\ \text{a1\_now}(X_0) \\ \text{named}(X_1\text{,maria,per}) \end{array}\right) \wedge \left(\begin{array}{|l} X_2\ X_3\ X_4 \\ \hline \text{n1\_sweater}(X_2) \\ \text{n1\_time}(X_4) \\ \text{n1\_event}(X_3) \\ \text{a1\_new}(X_2) \\ \text{a1\_blue}(X_2) \\ \text{v1\_buy}(X_3) \\ \text{r1\_agent}(X_3\text{,}X_1) \\ \text{r1\_patient}(X_3\text{,}X_2) \\ \text{r1\_temp\_included}(X_3\text{,}X_4) \\ \text{r1\_temp\_before}(X_4\text{,}X_0) \end{array}\right)$$

- A person named "Maria" exists. This person bought a new blue sweater.
- DRdoc-STM(Example_1$_a$, Example_1$_b$) = 18/71 = 0.2535.

Experimental Results

- Data: *'mt06'* part of the development set
  - Extracted from the NIST 2006 Open MT Evaluation Campaign Arabic-to-English translation exercise
  - 25 documents totalling 249 segments
  - 8 system and 4 reference translations
  - Adequacy assessments (average system adequacy is 5.38)
- Meta-evaluation in terms of *correlation coefficients with adequacy assessments*. We have produced document-level assessments by averging over individual segments.

Document-level Performance

Metric	Pearson$_\rho$	Spearman$_\rho$	Kendall$_\tau$
METEOR	0.9182	0.8478	0.6728
DR-$O_r(\star)$	0.8567	0.8061	0.6193
DR-$O_{rp}(\star)$	0.8286	0.7790	0.5875
DR-STM	0.7880	0.7468	0.5554
DR$_{doc}$-$O_r(\star)$	*0.7936*	*0.7784*	*0.5875*
DR$_{doc}$-$O_{rp}(\star)$	*0.7219*	*0.6737*	*0.4929*
DR$_{doc}$-STM	*0.7553*	*0.7421*	*0.5458*

- DR$_{doc}$ variants obtain *lower levels of correlation* than their DR counterparts.
- Possible explanation for this negative result:
  - Document-level quality assessments have been obtained by averaging segment-level assessments. This may be biasing the correlation.
  - Parsing errors could be causing the metric to confer less informed scores. This is especially relevant taking into account that candidate translations are not always well-formed.
  - Similarity measures employed may not be able to take advantage of the document-level features provided by the discourse analysis.

Conclusions and Future Steps

A proposal has been presented for document-level automatic MT evaluation based on Discourse Representations. Results presented are, however, preliminary. Further work must be conducted before reporting on the validity/applicability of our approach.

- Future Work
  1. Error Analysis
     - Analyse in deep detail the behavior of DR$_{doc}$ metrics
     - Study the impact of parsing errors
  2. Repeat this experiment over other test beds with document structure (WMT09, NIST09, ...)
  3. Explore the possibility of producing document-level assessments ourselves
  4. Refine the metric (identification and analysis of discourse markers)
  5. Define new metrics possibly using alternative linguistic processors.

References

[Bos08] Johan Bos. Wide-coverage semantic analysis with boxer. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Research in Computational Semantics, pages 277–286. College Publications, 2008.

[CCB07] James Curran, Stephen Clark, and Johan Bos. Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, 2007.

Free Download:[a] `http://www.lsi.upc.edu/~nlp/IQMT/`

[a] **Acknowledgements:** This research has been partially funded by the Spanish Government (projects OpenMT-2, TIN2009-14675-C03, and KNOW, TIN-2009-14715-C0403) and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement numbers 247762 (FAUST project, FP7-ICT-2009-4-247762) and 247914 (MOLTO project, FP7-ICT-2009-4-247914).