Learning to Differentiate Better from Worse Translations

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, Alessandro Moschitti, Preslav Nakov and **Massimo Nicosia**

Qatar Computing Research Institute

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

1. Task Formulation

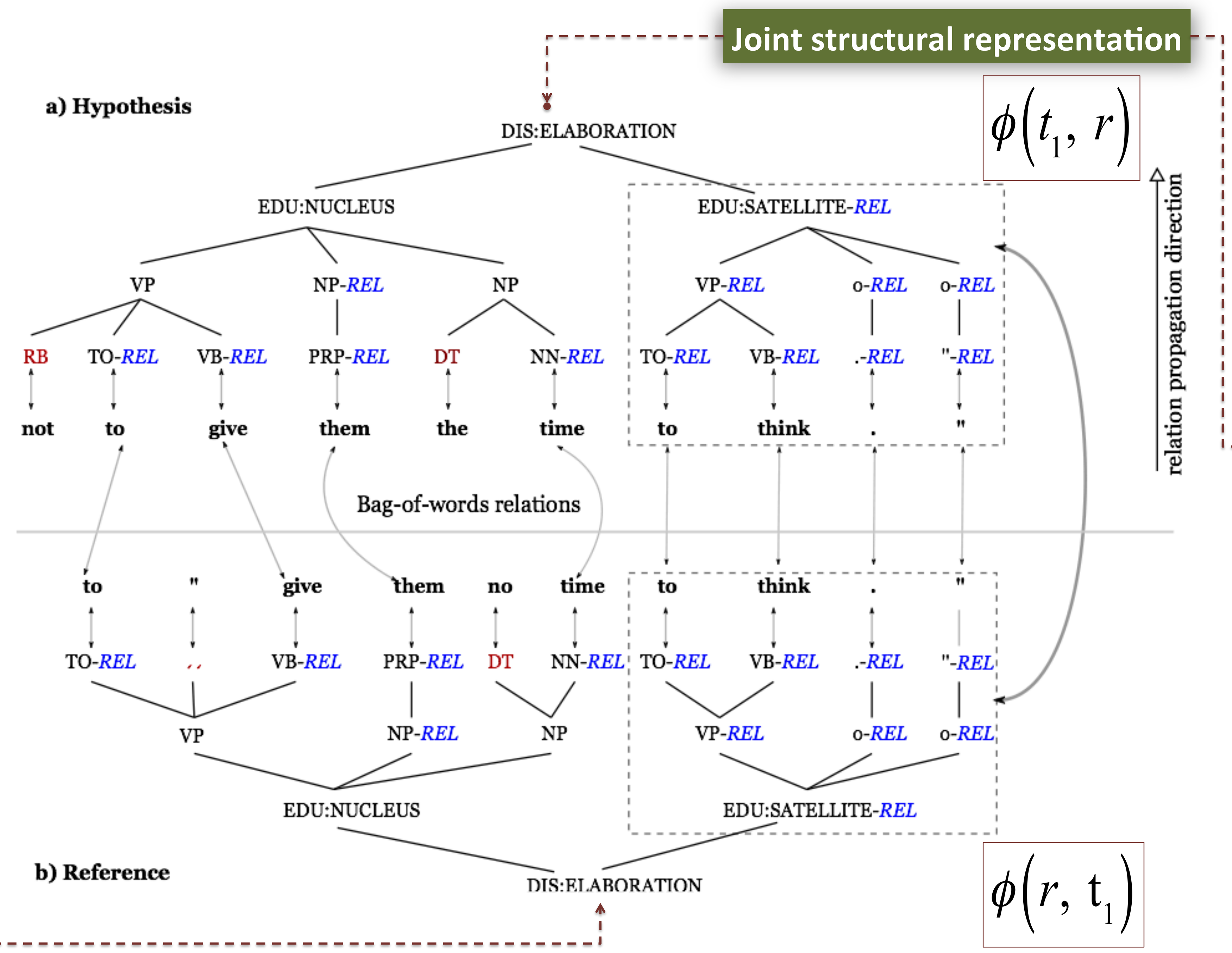
- Decide which of two alternative translations t_1 and t_2 is better given the reference r
- **Motivation**: Higher correlation with human judgments compared to absolute scores (Duh,2008; Song&Cohn,2011)

2. Proposed Solution

- Use the framework of **structured kernel learning** (Severyn&Moschitti, 2012)
  - Pairwise learning to rank formulation with kernels
  - Is more powerful than kernel similarity (Guzmán et al., 2014)
  - Learns features (structure fragments) automatically
  - Allows integrating several information sources

  Highlights!
- Integrate *lexical, syntactic,* and *discourse information* in a single *structural representation*
- Use both reference and system output simultaneously
  - **Learning object:** $\langle t_1, t_2, r \rangle$

3. Enriched Structural Representation



4. Structured Kernel Learning

Preference Kernel

$$PK(\langle t_1, t_2, r \rangle \langle t_1', t_2', r' \rangle) = K(t_1, t_1' | r, r') + K(t_2, t_2' | r, r') + K(t_1, t_2' | r, r') + K(t_2, t_1' | r, r')$$

Kernel

Where $K(t_1, t_1' | r, r') = PTK(\phi(t_1, r), \phi(t_1', r')) + PTK(\phi(r, t_1), \phi(r', t_1'))$

Partial Tree Kernel

5. Experimental Settings

- **Train**: 10K judgments per language (WMT-11)
- **Langs**: Czech–English (cs-en), German-English (de-en), Spanish-English (es-en), French-English (fr-en)
- **Eval**: Kendall's Tau as a measure of correlation on WMT-12 data (official)
- **Results are compared with direct kernel similarity**

6. Evaluation Results

Train & Test for each language pair separately on different structures

		Similarity					Structured Kernel Learning				
	Structure	*cs*-en	*de*-en	*es*-en	*fr*-en	all	*cs*-en	*de*-en	*es*-en	*fr*-en	all
1	SYN	0.169	0.188	0.203	0.222	0.195	0.190	0.244	0.198	0.158	0.198
2	DIS	0.130	0.174	0.188	0.169	0.165	0.176	0.235	0.166	0.160	0.184
3	DIS+POS	0.135	0.186	0.190	0.178	0.172	0.167	0.232	0.202	0.133	0.183
4	DIS+SYN	0.156	0.205	0.206	0.203	0.192	**0.210**	**0.251**	**0.240**	**0.223**	**0.231**

SYN (syntactic parse), DIS (RST discourse parse relations), POS (part of speech)

Observations

- **Learning with structural kernels works better than using simple kernel similarity ⇒ new features are learned**
- Shallow syntax and discourse yield similar improvement individually
- Combining them yields further improvement
- **We outperform popular metrics like TER (0.217), NIST (0.214) and BLEU (0.185)**

Cross-language training and testing

		Testing				
	Train	*cs*-en	*de*-en	*es*-en	*fr*-en	all
1	*cs*-en	0.210	0.204	0.217	0.204	0.209
2	*de*-en	0.196	0.251	0.203	0.202	0.213
3	*es*-en	0.218	0.204	**0.240**	0.223	0.221
4	*fr*-en	0.203	0.218	0.224	0.223	0.217
5	all	**0.231**	**0.258**	0.226	**0.232**	**0.237**

Observations

- Same language training is better in most cases
- However, overall differences are rather small
- Training on all language pairs yields the best results in all cases except for es-en

7. Conclusion

- Unified framework for integrating layers of linguistic information for MT evaluation
- Pairwise learning-to-rank with structural kernels
- Competitive performance

8. Future Work

- More linguistic information: SRL, Brown clusters, etc.
- Integrate scores from other MT evaluation metrics
- Use of more relations between *t* and *r*.

Acknowledgments

This research is part of the Interactive sYstems for Answer Search (Iyas) project, conducted by the Arabic Language Technologies (ALT) group at Qatar Computing Research Institute (QCRI) within the Qatar Foundation