DiscoTK: Using Discourse Structure for Machine Translation Evaluation

Shafiq Joty, Francisco Guzmán, Lluís Màrquez and **Preslav Nakov**
Qatar Computing Research Institute

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

Discourse for MT Evaluation

- **Discourse structure helps MT evaluation** (Guzmán et al., 2014)
- We present two metrics that consider discourse information
  - DiscoTK$_{light}$ only uses discourse
  - DiscoTK$_{party}$ also uses metrics from ASIYA
- **DiscoTK$_{party}$ is the best performing metric at WMT14**

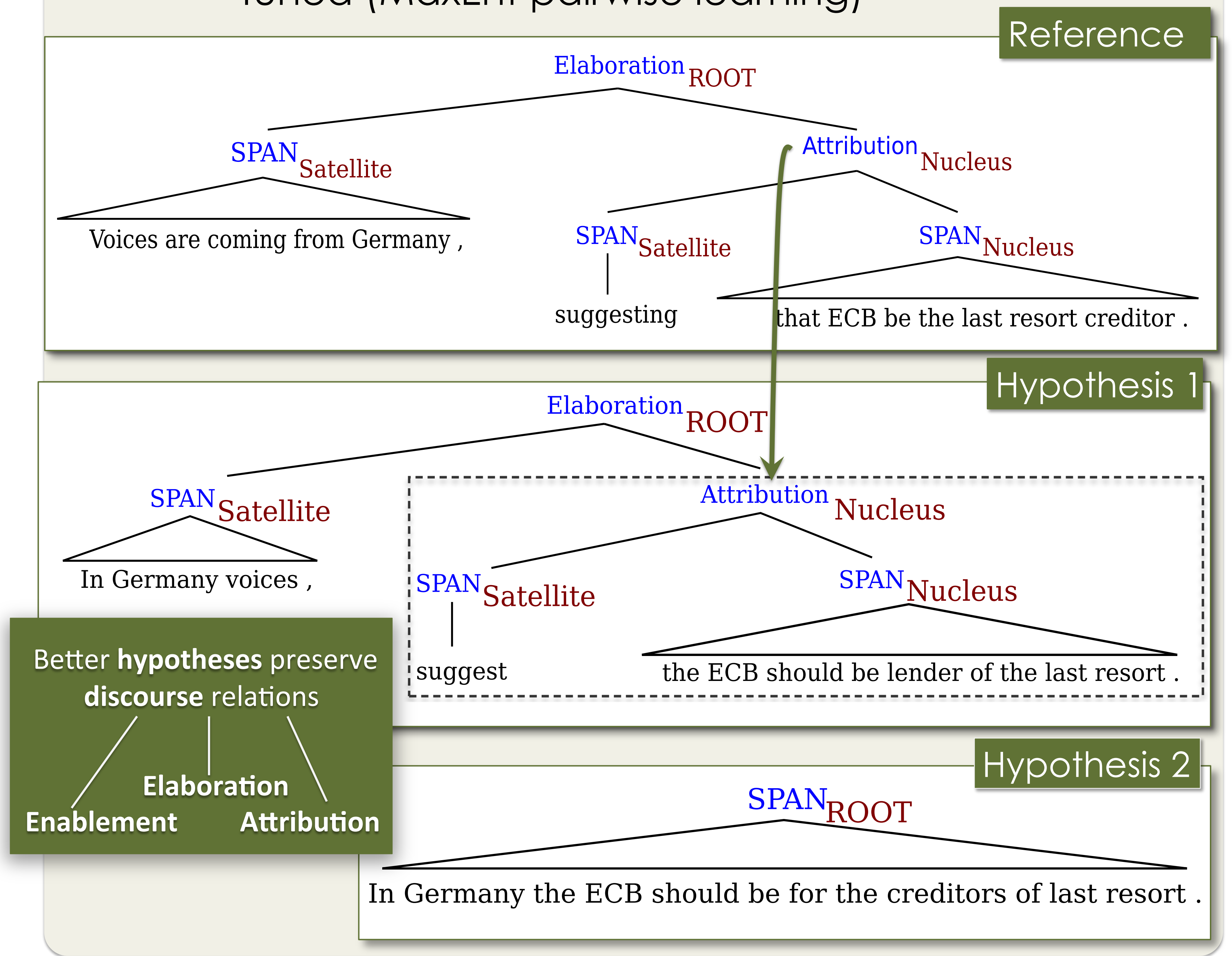
Method

Compute discourse similarity between *Hyp* and *Ref*

- RST-parse *Hyp* and *Ref* (Joty et al., 2012)
- RST trees are transformed to five different representations
- We use syntactic tree kernel (Collins & Duffy, 2002) to measure the similarity between two discourse trees
  - Use this similarity as a segment-level score
  - For system-level, average segment level scores

Combine discourse similarity with existing metrics (ASIYA)

- Uniform linear interpolation
- Tuned (MaxEnt pairwise learning)

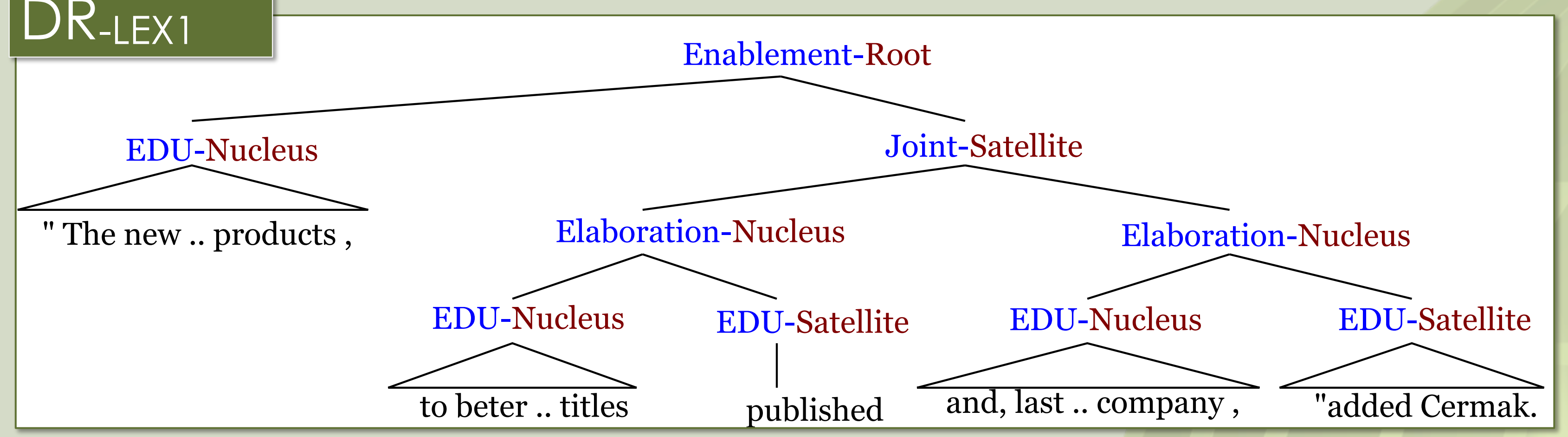


Discourse-based Metrics

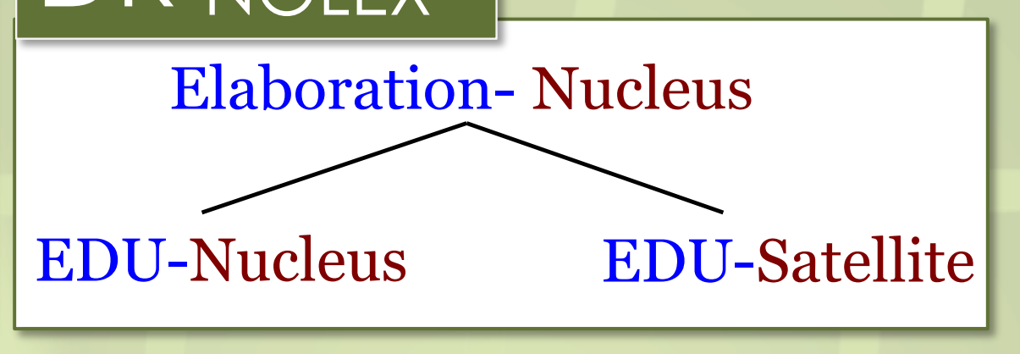
DiscoTK$_{party}$

DiscoTK$_{light}$

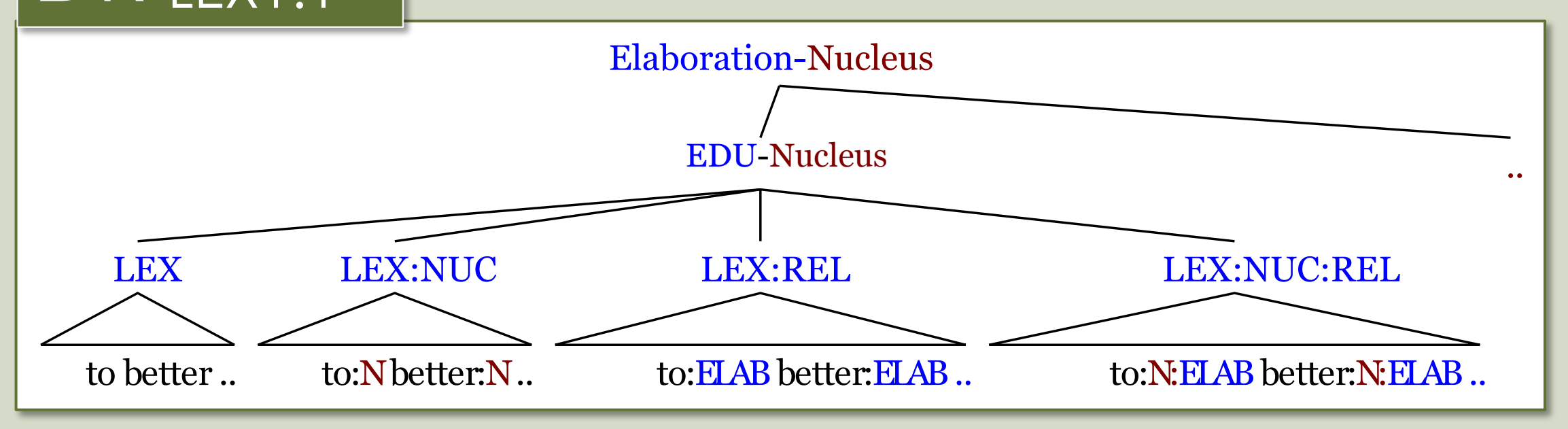
DR-LEX1



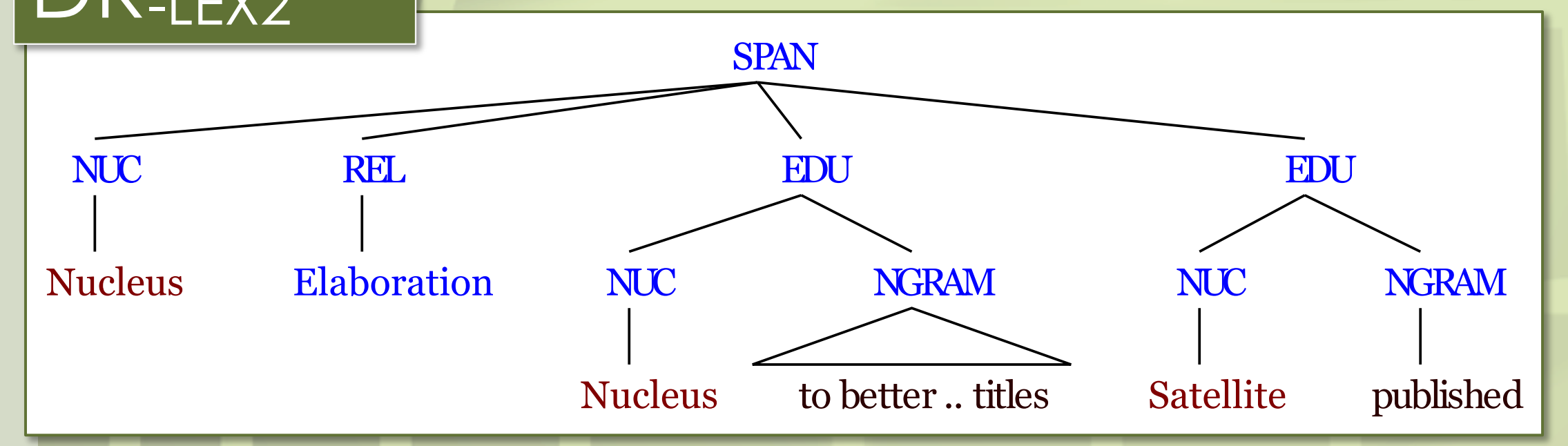
DR-NOLEX



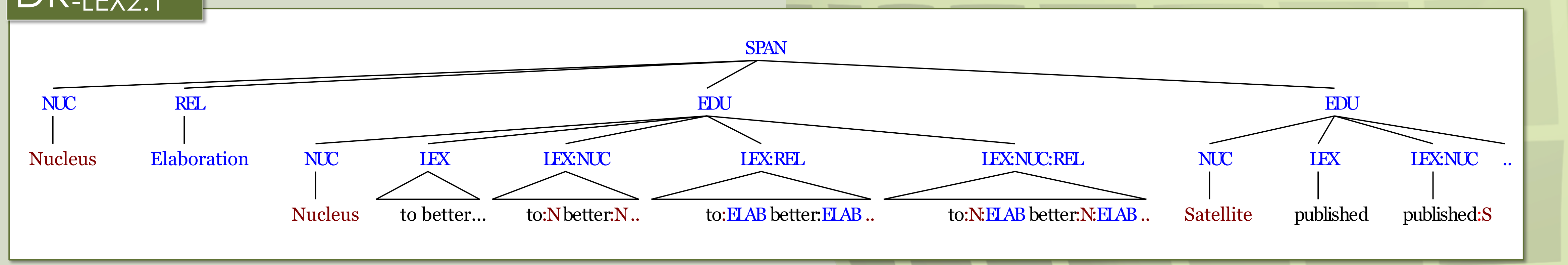
DR-LEX1.1



DR-LEX2



DR-LEX2.1



+

ASIYA

BLEU
NIST
TER
TERp-A
ROUGE-W
METEOR-ex
METEOR-pa
METEOR-st
METEOR-sy
DP-HWCM_c-4
DP-HWCM_r-4
DP-Or(*)
CP-STM-4
SR-Or(*)
SR-Mr(*)
SR-Or
DR-Or(*)
DR-Orp(*)

Results

Metric	Tuning	Segment Level WMT12 τ	Segment Level WMT13 τ	System Level WMT12 ρ	System Level WMT12 r	System Level WMT13 ρ	System Level WMT13 r
SEMPOS	na	–	–	0.902	0.922	–	–
SPEDE07PP	na	0.254	–	–	–	–	–
METEOR-WMT13	na	–	0.264	–	–	0.935	**0.950**
DiscoTK$_{light}$	∅	0.171	0.162	0.884	0.922	0.880	0.911
	WMT11	0.207	0.201	0.860	0.872	0.890	0.909
	WMT12	–	0.200	–	–	0.889	0.910
	WMT13	0.206	–	0.865	0.871	–	–
	WMT11+12	–	0.197	–	–	0.890	0.910
	WMT11+13	0.207	–	0.865	0.871	–	–
DiscoTK$_{party}$	∅	0.257	0.231	0.907	0.915	**0.941**	0.928
	WMT11	0.302	0.282	**0.915**	**0.940**	0.934	0.946
	WMT12	–	0.284	–	–	0.936	0.940
	WMT13	**0.305**	–	0.912	0.935	–	–
	WMT11+12	–	**0.289**	–	–	0.936	0.943
	WMT11+13	0.304	–	0.912	0.934	–	–
ASIYA	∅	0.273	0.252	0.899	0.909	0.932	0.922
	WMT11	0.301	0.279	0.913	0.935	0.934	0.944
	WMT12	–	0.277	–	–	0.932	0.938
	WMT13	0.303	–	0.908	0.932	–	–
	WMT11+12	–	0.277	–	–	0.934	0.940
	WMT11+13	0.303	–	0.908	0.933	–	–

Summary

- DiscoTK$_{light}$ competitive at system-level
- Tuned DiscoTK$_{party}$ improves over ASIYA both at segment- and system-level
- Tuning helps consistently
- We improve over the best WMT12, WMT13 results

Tuned DiscoTK$_{party}$ ranked 1st at WMT14

Future Work

- Learn with preference kernels from a syntactic-semantic-discourse tree representation
- Go beyond the sentence-level