Linguistically-Enriched Models for Bulgarian-to-English Machine Translation

Rui Wang
DFKI GmbH, Germany
(collaboration with Petya Osenova
and Kiril Simov, BAS-IICT, Bulgaria)

In a Nutshell

- Bulgarian → English
- Factored SMT models to incorporate linguistic knowledge
- Question-based manual evaluation

Motivation

- Incorporating linguistic knowledge into statistical models, same for MT

- Different strategies
  - Post-editing
  - System combination

Our Strategy

- Good baseline result (38.61 BLEU by Moses)
- Various linguistic knowledge from preprocessing
  - Morphological analysis, lemmatization, POS tagging
  - (CoNLL) Syntactic dependency tree
  - (R)MRS
- 'Supertagging'-style

Related Work

- Birch et al. (2007) and Hassan et al. (2007)
  - Supertags on English side
- Singh and Bandyopadhyay (2010)
  - Manipuri-English bidirectional translation
- Bond et al. (2005), Oepen et al. (2007), Graham and van Genabith (2008), and Graham et al. (2009)
  - Transfer-based MT

Factored Model

- Koehn and Hoang (2007)
  - Easily incorporate linguistic features at the token level
  - Similar to 'supertags'

- WF, Lemma, POS, Ling
- DepRel, HLemma, HPOS
- EP, EoV, ARGnEP, ARGnPOS

Preprocessing

- POS Tagging – 97.98% accuracy
  Lemmatization – 95.23 % accuracy
  - Georgiev et al., 2012

- Dependency Parsing – 87.6 % labeled parsing accuracy
  - Savkov et al., 2012

Example

- *Spored odita v elektricheskite kompanii politicite zloupotrebyavat s dyrzhavnite predpriyatiya.*

- *Electricity audits prove politicians abusing public companies.*

Factors

No	WF	Lemma	POS	Ling	DepRel	HLemma	HPOS
1	spored	spored	R	_	adjunct	zloupotrebyavam	VP
2	odita	odit	Nc	npd	prepcomp	spored	R
3	v	v	R	_	mod	odit	Nc
4	elektricheskite	elektricheski	A	pd	mod	kompaniya	Nc
5	kompanii	kompaniya	Nc	fpi	prepcomp	v	R
6	politicite	politik	Nc	mpd	subj	zloupotrebyavam	Vp
7	zloupotrebyavat	zloupotrebyavam	Vp	tir3p	root	-	-
8	s	s	R	_	indobj	zloupotrebyavam	Vp
9	dyrzhavnite	dyrzhaven	A	pd	mod	predpriyatie	Nc
10	predpriyatiya	predpriyatie	Nc	npi	prepcomp	s	R

Minimal Recursion Semantics (MRS)

- MRS Structure: <GT, R, C>
  - GT: *Top*
  - R: a bag of EPs
  - C: *Handle Constraints*, the outscopes order between the EPs

- Examples:
  - *<h0, {h1:every(x, h2, h3), h2:dog(x), h4:chase(x, y), h5:some(y, h6, h7), h6:white(y), h6:cat(y)}, {}>*

(Fallback) Rules for RMRS

- <Lemma, MSTag> → EP-RMRS
  - The rules of this type produce an RMRS including an elementary predicate

- <DRMRS, Rel, HRMRS> → HRMRS'
  - The rules of this type unite the RMRS constructed for a dependent node (DRMRS) into the current RMRS for a head node (HRMRS)

Factors (cont.)

No	EP	EoV	EP_1/POS_1	EP_2/POS_2	EP_3/POS_3
1	spored_r	e	zloupotrebyavam_v/Vp	odit_n/Nc	-
2	odit_n	v	-	-	-
3	v_r	e	odit_n/Nc	kompaniya_n/Nc	-
4	elekticheski_a	e	kompaniya_n/Nc	-	-
5	kompaniya_n	v	-	-	-
6	politik_n	v	-	-	-
7	zloupotrebyavam_v	e	politik_n/Nc	-	s_r/R
8	s_r	e	zloupotrebyavam_v/Vp	predpriyatie_n/Nc	-
9	dyrzhaven_a	e	predpriyatie_n/Nc	-	-
10	predpriyatie_n	v	-	-	-

Example

(1) Momcheto j go dava buketa na
Boy-the her-dat it-acc gives bouquet-the to
momicheto.
girl-the.
The boy gives the bouquet to the girl.

(2) Momcheto j go dava.
Boy-the her-dat it-acc gives.
The boy gives it to her.

Experiments

- GIZA++ (Och and Ney, 2003)
- A tri-gram language model is estimated using the SRILM toolkit (Stolcke, 2002)
- Minimum error rate training (MERT) (Och, 2003) is applied to tune the weights for the set of feature weights that maximizes the BLEU score on the development se

Corpora

- Train/Dev/Test

- SETIMES
  - 150,000(100,000)/500/1,000
- EMEA
  - 700,000/500/1,000
- JRC-Acquis
  - 0/0/4,107

Results

Corpora	Test	Dev	Final	Drop
SETIMES → SETIMES	34.69	37.82	36.49	/
EMEA → EMEA	51.75	54.77	51.62	/
SETIMES → EMEA	13.37	/	/	61.5%
SETIMES → JRC-Acquis	7.19	/	/	79.3%
EMEA → SETIMES	7.37	/	/	85.8%
EMEA → JRC-Acquis	9.21	/	/	82.2%

Results (cont.)

ID	Model	BLEU	1-gram	2-gram	3-gram	4-gram	METEOR
1	WF (Baseline)	38.61	**69.9**	44.6	31.5	22.7	0.3816
2	WF, POS	**38.85**	**69.9**	**44.8**	**31.7**	**23.0**	0.3812
3	WF, LEMMA, POS, LING	38.84	**69.9**	44.7	**31.7**	**23.0**	0.3803
4	LEMMA	37.22	68.8	43.0	30.1	21.5	**0.3817**
5	LEMMA, POS	37.49	68.9	43.2	30.4	21.8	0.3812
6	LEMMA, POS, LING	38.70	69.7	44.6	31.6	22.8	0.3800
7	WF, DEPREL	36.87	68.4	42.8	29.9	21.1	0.3627
8	WF, DEPREL, HPOS	36.21	67.6	42.1	29.3	20.7	0.3524
9	WF, LEMMA, POS, LING, DEPREL	36.97	68.2	42.9	30.0	21.3	0.3610
10	WF, POS, EP	38.74	69.8	44.6	31.6	22.9	0.3807
11	WF, EP, EOV	38.74	69.8	44.6	31.6	22.9	0.3807
12	WF, POS, LING, EP, EOV	38.76	69.8	44.6	**31.7**	22.9	0.3802
13	EP, EOV	37.22	68.5	42.9	30.2	21.6	0.3711
14	EP, EOV, LING	38.38	69.3	44.2	31.3	22.7	0.3691
15	EP, EOV, ARG_nPOS	36.21	67.4	41.9	29.2	20.9	0.3577
16	WF, EP, EOV, ARG_nPOS	37.37	68.4	43.2	30.3	21.8	0.3641

Manual Evaluation

- Motivation
  - BLEU score in high range is not differentiable
  - Impacts from various linguistic knowledge
- Evaluation metrics
  - Grammaticality
  - Content

Results

ID	Model	Grammaticality						Content						Final
		1	2	3	4	5	Sum	1	2	3	4	5	Sum	
1	WF (Baseline)	20	47	5	32	**46**	3.25	20	46	5	23	56	3.33	3.29
2	WF, POS	20	48	5	37	40	3.19	20	48	5	24	53	3.28	3.24
3	WF, LEMMA, POS, LING	20	47	6	34	43	3.22	20	47	1	24	58	3.35	3.29
4	LEMMA	**15**	34	11	46	44	**3.47**	15	32	5	**33**	**65**	**3.67**	**3.57**
5	LEMMA, POS	**15**	38	12	**51**	34	**3.34**	15	35	9	32	59	**3.57**	**3.45**
6	LEMMA, POS, LING	20	48	5	34	43	3.21	20	48	5	22	55	3.29	3.25
7	WF, DEPREL	32	48	3	29	38	2.95	32	49	4	14	51	3.02	2.99
8	WF, DEPREL, HPOS	45	41	7	23	34	2.73	45	41	2	21	41	2.81	2.77
9	WF, LEMMA, POS, LING, DEPREL	34	47	5	30	34	2.89	34	48	3	20	45	2.96	2.92
10	WF, POS, EP	19	49	4	34	44	3.23	19	49	3	20	59	3.34	3.29
11	WF, EP, EOV	20	49	2	41	38	3.19	19	50	4	16	61	3.33	3.26
12	WF, POS, LING, EP, EOV	19	49	5	37	40	3.20	19	50	3	24	54	3.29	3.25
13	EP, EOV	**15**	41	10	44	40	**3.35**	**14**	38	7	31	60	**3.57**	**3.46**
14	EP, EOV, LING	20	49	7	38	36	3.14	19	49	7	20	55	3.29	3.21
15	EP, EOV, ARG$_n$POS	23	49	9	34	35	3.06	23	47	8	**33**	39	3.12	3.09
16	WF, EP, EOV, ARG$_n$POS	34	47	10	30	29	2.82	34	47	10	20	39	2.89	2.85
*	GOOGLE	0	2	20	52	76	4.35	1	0	9	42	98	4.57	4.46
*	REFERENCE	0	0	5	51	94	4.59	1	0	5	37	107	4.66	4.63

Question-Based Evaluation

- Either like it or dislike it
- A set of questions based on dependency relations
- Answers to judge
- Similar to PETE (Yuret te al., 2010)

ID	Model	Score
1	WF (Baseline)	127
2	WF, POS	126
3	WF, LEMMA, POS, LING	131
4	LEMMA	**133**
5	LEMMA, POS	**133**
6	LEMMA, POS, LING	128
7	WF, DEPREL	131
8	WF, DEPREL, HPOS	120
9	WF, LEMMA, POS, LING, DEPREL	124
10	WF, POS, EP	125
11	WF, EP, EOV	126
12	WF, POS, LING, EP, EOV	128
13	EP, EOV	**138**
14	EP, EOV, LING	122
15	EP, EOV, ARG$_n$POS	130
16	WF, EP, EOV, ARG$_n$POS	121

Conclusion

- Factored model is nice tool to incorporate morphological features
  - Sparsity
- Syntactic/Semantic information without *structure* is not so helpful
  - Deeper transfer

More Issues

- Morphology
  - Somehow handled by the factored model
- Semantic empty words
  - Difficult for word alignment
- Reordering
  - Difficult without structural information

Acknowledgements

- EuroMatrixPlus (IST-231720)

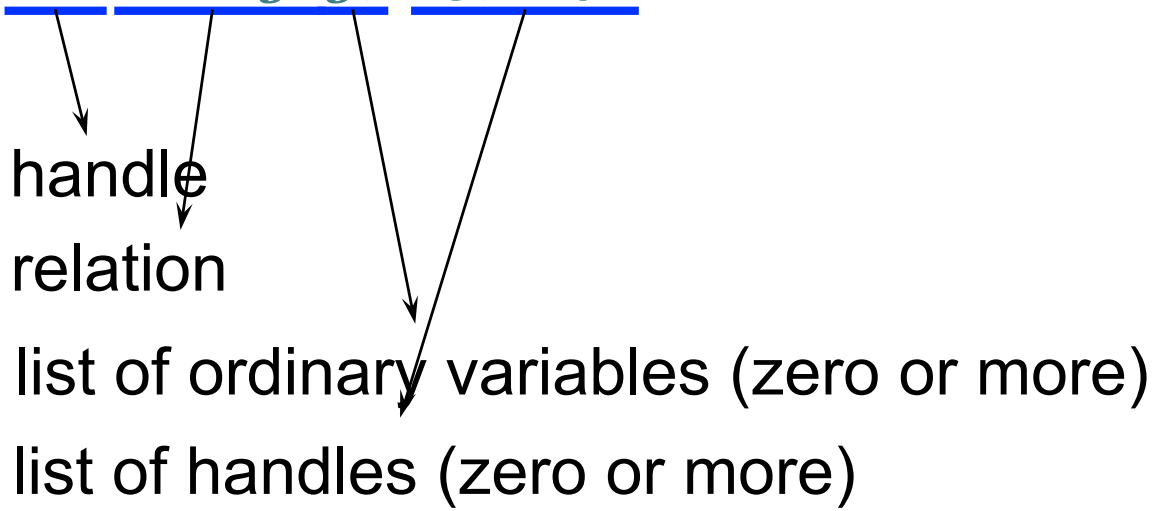
- Tania Avgustinova for fruitful discussions and her helpful linguistic analysis
- Laska Laskova, Stanislava Kancheva and Ivaylo Radev for doing the human evaluation of the data

Thank YOU!

Questions?

MRS (cont.)

- Elementary Predication (EP)
  - *h2:every(y, h3, h4)*

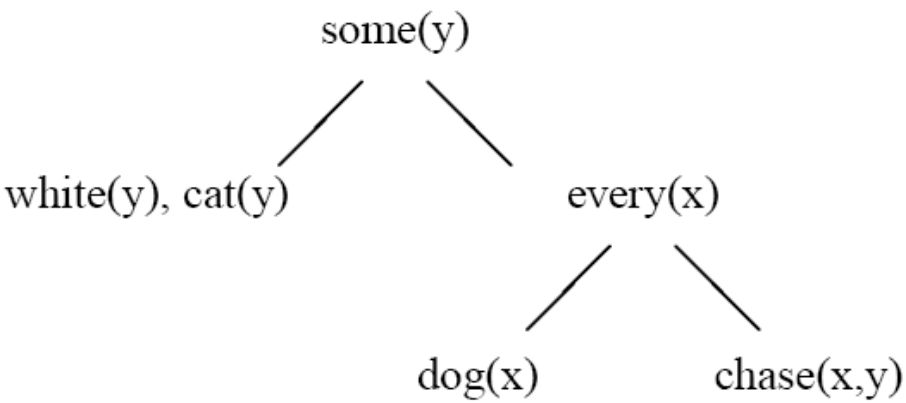


Scope Underspecification

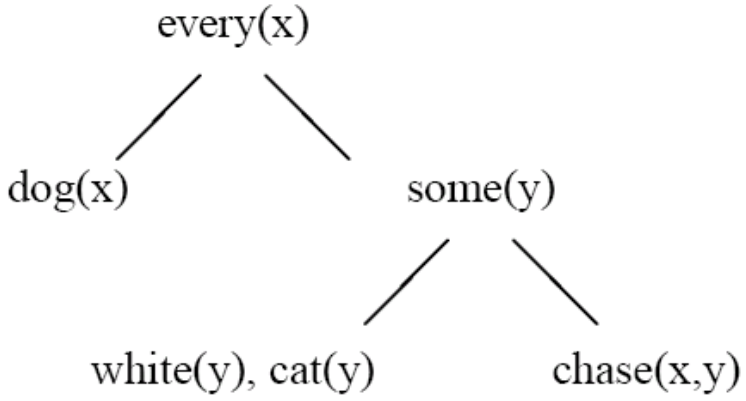
- Examples
  - *Every dog chases some white cat.*

(a) some(y, white(y)∧cat(y), every(x, dog(x), chase(x, y)))

(b) every(x, dog(x), some(y, white(y)∧cat(y), chase(x, y)))



h1:every(x, h3, h4), h3:dog(x),
h7:white(y), h7:cat(y),
h5:some(y, h7, h1), h4:chase(x,y)



h1:every(x, h3, h5), h3:dog(x),
h7:white(y), h7:cat(y),
h5:some(y, h7, h4), h4:chase(x, y)

Manual Evaluation – Grammaticality

1. The translation is not understandable.
2. The evaluator can somehow guess the meaning, but cannot fully understand the whole text.
3. The translation is understandable, but with some efforts.
4. The translation is quite fluent with some minor mistakes or re-ordering of the words.
5. The translation is perfectly readable and grammatical.

Manual Evaluation – Content

1. The translation is totally different from the reference.
2. About 20% of the content is translated, missing the major content/topic.
3. About 50% of the content is translated, with some missing parts.
4. About 80% of the content is translated, missing only minor things.
5. All the content is translated.