Grammar Knowledge Transfer for Building RMRSs over Dependency Parses in Bulgarian

Kiril Simov and **Petya Osenova**

Linguistic Modelling Department, IICT,

Bulgarian Academy of Sciences

DELPH-IN, Sofia, 2012

Plan of the Talk

- Goal
- Related Work
- RMRS analysis
- Conclusions and Future Work

Goal

- Create a pipeline for RMRS analysis of Bulgarian
- We rely on the Bulgarian HPSG resource grammar BURGER, and on a dependency parser (Malt Parser – Nivre et al. 2006), trained on the BulTreeBank data

Related Work

Our work is inspired by:

- (Copestake, 2003; 2007) on MRS and RMRS
- The previous work on transfer of dependency analyses into RMRS structures described in:
  - (Spreyer and Frank, 2005) for TIGER treebank of German, and
  - (Jakob et al, 2010) Prague Dependency Treebank of Czech (PDT)

RMRS

In the paper we follow the representation of RMRS used in (Jakob et al, 2010), which defines an RMRS structure as a quadruple:

< *hook*, *EP-bag*, *argument set*, *handle constraints* >

Bulgarian Language Pipeline

- BURGER – Bulgarian Resource Grammar
- Language Processing Pipeline:
  - Tokenization and sentence boundary identification
  - POS tagging with 97.98 % accuracy (680 tags)
  - Lemmatization with 95.23 % accuracy
  - Dependency Parsing with 87.6 % labeled parsing accuracy (17 tags)
  - RMRS analysis over dependency parsing

Bulgarian Dependency Tagset

adjunct 12009	Adjunct (optional verbal argument)	**subj** 14064	Subject
clitic 2263	Short forms of the possessive pronouns	**pragadjunct** 1612	Pragmatic adjunct
comp 18043	Complement (argument of non-finite verbs, copula, auxiliaries)	**punct** 28134	Punctuation
conj 6342	Conjunction in coordination	**xadjunct** 1826	Clausal adjunct
conjarg 7005	Argument (second, third, ...) of coordination	**xcomp** 4651	Clausal complement
indobj 4232	Indirect Object	**xmod** 2219	Clausal modifier
marked 2650	Marked (clause, introduced by a subordinator)	**xprepcomp** 168	Clausal complement of preposition
mod 42706	Modifier	**xsubj** 504	Clausal subject
obj 7248	Object (direct argument of a non-auxiliary verbal head)		

Input for RMRS

The information for the RMRS structures is based on the following linguistic annotation:

- the lemma (*Lemma*) for the given wordform;
- the morphosyntactic tag (*MSTag*) of the wordform, and
- the dependent relations in the dependency tree
- In cases of quantifiers we have access to the lexicon used in BURGER

Rules for RMRS

- Two types:
  - *<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*)

Examples: verb чета (‘*read-I’, I read*)

Rule:

<Lemma, Vp> →

< l1:a1:e1, { l1:a1:lemma_v_rel(e1) }, { a1:ARG1(x1) }, {} >

… чета (**чета**, **Vp**) ...

< l1:a1:e1, { l1:a1:чета_v_rel(e1) }, { a1:ARG1(x1) }, {} >

Examples: verb чета му я (‘*read him her’* I read it to him) (1)

Rule:

< <l2:a2:x2, {}, {a2:ARG2(x2)}, HC1>, comp,

< l1:a1:e1, { l1:a1:lemma_v_rel(e1) | R }, ARGS, HC2 > >

→

< l1:a1:e1, { l1:a1:lemma_v_rel(e1) | R },

{ a1:ARG2(x2) } ∪ ARGS, HC1 ∪ HC2 >

Examples: verb чета му я (‘*read him her’* I read it to him) (2)

<l2:a2:x2, {}, {a2:ARG2(x2)}, {}>,

comp,

< l1:a1:e1, { l1:a1:чета_v_rel(e1) }, { a1:ARG1(x1) }, {} >

→

< l1:a1:e1,

{ l1:a1:чета_v_rel(e1) },

{ a1:ARG1(x1), a1:ARG2(x2) },

{} >

Examples: verb чета му я (‘*read him her’* I read it to him) (3)

<l3:a3:x3, {}, {a3:ARG2(x3)}, {}>

is incorporated in a similar way:

< l1:a1:e1,
{ l1:a1:чета_v_rel(e1) },
{ a1:ARG1(x1), a1:ARG2(x2), a1:ARG3(x3) },
{} >

момче му я чете (Boy him-dative her-accusative read, ‘A boy reads it to him’)

Rule:

< <l2:a2:x2, {l2:a2:lemma_n_rel(x1)|R1}, ARGS2, HC2>, subj,
< l1:a1:e1, { l1:a1:lemma_v_rel(e1) | R2 }, ARGS1, HC1 > >

→

< l1:a1:e1, { l1:a1:lemma_v_rel(e1), l2:a2:lemma_n_rel(x1) |
R1 ∪ R2 },
{ a1:ARG1(x1) } ∪ ARGS1 ∪ ARGS2 , HC1 ∪ HC2 >

момче му я чете (Boy him-dative her-accusative read, ‘A boy reads it to him’)

< ‘момче’ > subj <‘му я чете’ >

→

< l2:a4:e1,

{ l1:a1:момче_n_rel(x1), l2:a4:чета_v_rel(e1) },

{ a4:ARG1(x1), a4:ARG2(x2), a4:ARG3(x3) },

{} >

момче му чете книга (Boy him-dative reads book, ‘A boy reads a book to him’

< ‘момче’ > subj <‘му чете’ > and

< ‘книга’ > obj <‘му чете’ >

→

< l2:a3:e1,

{ l1:a1:момче_n_rel(x1), l2:a3:чета_v_rel(e1),

l3:a4:книга_n_rel(x2) },

{ a3:ARG1(x1), a3:ARG2(x2), a3:ARG3(x3) }, {} >

Algorithm

The dependency tree is traversed two times:

1. Top-down for each lexical node the RMRS on the basis of lemma and morphosyntactic information is constructed
2. Then bottom-up the RMRS for the dependent elements are incorporated within the head RMRS

Conclusions and Future Work

- We have developed a pipeline which produces RMRS analysis for Bulgarian sentences
- We have exploited it in Bulgarian-English SMT
- Improving the RMRS details and accuracy
- Using RMRS analysis to support the deep analysis with BURGER