Word Sense Annotation, Disambiguation and Knowledge Transfer for Bulgarian

Kiril Simov and Petya Osenova

Linguistic Modelling Department
Institute for Information and Communication Technology
Bulgarian Academy of Sciences
AComIn Project

29 July 2015
CLTL, Amsterdam, the Netherlands
Plan of the Talk

• Brief update on our work
• Sense annotation
• SRL
• WSD experiments
• Conclusions
IICT – VU Exchange Visit

This scientific visit is supported by the EU Project AComIn: Advanced Computing for Innovation, grant N 316087: http://www.iict.bas.bg/acomin/
AComIn: Advanced Computing in Innovation

- A 3,2 M€ grant in FP7 Capacity with a single beneficiary – IICT-BAS

- Objectives:
  - Strengthening Human Potential
  - Providing up-to-date Research Infrastructure
  - Focus on users
  - Networking with EU partners
  - Strengthening the IICT-BAS Innovation Capacity
  - Dissemination via various events/channels
  - Organising assessment of the IICT-BAS achievements
Last Year – July 2014 - Amsterdam

• The Core Wordnet for Bulgarian was released via Open Multilingual Wordnet site: http://compling.hss.ntu.edu.sg/omw/

• Our NLP pipe was tuned to produce NAF output. The pipe for Bulgarian was included into NewsReader website.

• Preliminary ideas on Semantic Role Labeling. However, we did not have the treebank semantic annotation finished yet
Meanwhile: from July 2014 to July 2015

• Sense annotation of the treebank and WSD experiments with the UKB tool
• Extending our Wordnet with senses from the treebank
• Catena approach to MWEs and valency
• Using DBPedia for transliterating/translating foreign names into Bulgarian
• Transferring BulTreeBank into Universal Dependencies format (first release was on 15 May 2015 with 125k, which is half of the resource):
http://universaldependencies.github.io/docs/#language-bg
This year – July 2015 - Amsterdam

- Preparation for new WSD experiments with the UKB tool: extraction of relations from SemCor
- Transferring the predicate matrix information from the annotated English part of a news corpus (Setimes) to the Bulgarian part
- Cleaning the extended Wordnet
- Evaluation of the results for Bulgarian from Antoni’s WN tool
Sense Annotation

Two stages:

– Stage 1 - DONE
  • Mapping the definitions of a Bulgarian explanatory dictionary to the intersected senses of Core and Base Concepts in Princeton WordNet
  • Mappings manually checked and curated wrt: selection of the correct sense; addition of a sense or update of a definition

– Stage 2
  • Mapping nouns, verbs, adjectives and adverbs from the treebank to WordNet – ALMOST DONE
  • Annotation of domain specific texts (IT) with WordNet – STARTED
  • Using Antoni’s WN tool for extending the WordNet - STARTED
Sense Annotation: Process

Three layers:

– Verb valency frames [Osenova et. al. 2012]

– Senses of verbs, nouns, adjectives and adverbs

– DBPedia URIs over named entities.
Mapping of Treebank Senses to Wordnet

Partial concept correspondence dominates - the concepts differ in terms of specificity;
The use attribute \textit{rel(ation)} with three values – 0, 1, 2

- 0 – one-to-one correspondence (\textit{equality});
- 1 – a more specific definition in Bulgarian is mapped to a more general English definition;
- 2 – a more general definition in Bulgarian is mapped to a more specific English definition;
Synset Mapping

en-ss-01

en-ss-02

en-ss-03

en-ss-04

en-ss-05

en-ss-06

en-ss-07

subsumption relation

specificity relation

bg-def
DBPedia Linking

• DBPedia URIs annotation was performed as a separate activity
• It covered 10 885 named entities
• Unfortunately, the coverage of the Bulgarian DBPedia is rather small
• For that reason, the Bulgarian Wikipedia was explored
Sense Annotation: MWEs

• During sense annotation all the idiomatic expressions (*idioms, light verb constructions*, etc.) have been specifically labeled as multiword expressions (in contrast to the previously pure syntactic approach, taken in the annotation of the treebank)

• Since many of these expressions have a rather narrow potential for combination with other units, the differences show in the ontological constraints.

• Example: PERSON/GROUP OF PERSONS remains/remain without *roof* (PERSON/GROUP OF PERSONS becomes/become homeless)
Some Statistics

• Sense coverage of the BulTreeBank with Wordnet mappings:
  – 79,703 tokens;
  – 37,330 nouns;
  – 14,341 verbs;
  – 17,304 adjectives;
  – 10,728 adverbs.

All tokens: 107,961

• Verb valency frames:
  – 1755 verbs;
  – 3435 valency frames

• Extending WN with Antoni’s tool: after manual evaluation the precision was improved from 66.84% (only Core WN) to 76.53% (plus evaluated additions).
Some Observations on the PM transfer on Setimes Data

• Reasons for non-transfer
  – Error is POS tagger or lemma in BG
  – BG lemma is not present in WN

• Transferred cases
  – The sense differs from PM, but still holds
  – Several senses matched, and the correct one there
  – One sense matched from BG WN, but the correct one
  – Wrong sense match due to the missing BG sense in WN
The Sense Differs from PM, but still Holds

• Prodi sought to confirm
• Проди искаше (wanted) да се убеди

In EN: seek, hunt, look for (verb.contact))
In BG: want=desire (verb.emotion)
Several Senses Matched, Including the Correct one

- Ministers will not have summer holidays
- Много министерства няма да разрешават (allow) отпуск

Since two different verbs have been used: have and allow, the transferred concept from PM of Permission is correct
One Sense Matched from BG WN, but the Correct One

• This is expected in November
• Това се очаква през ноември

In BG WN there is only one mapping:
verb.cognition: regard something as probable or likely
Wrong Sense Match due to Missing BG

I am here to share (communication) the emotion

Аз съм тук, за да споделя (possession) вашите емоции

He must make (verb.social=carry out) a political decision

Той трябва да вземе (take=verb.possession) политическо решение
Knowledge-based WSD

• Our own toy implementation of Page Rank and Personalized Page Rank
• Small knowledge graph: ontology and relations
• Experiments with inheritance, structure of the graph, mappings from the text to the graph
• Easy to control and easy to observe the performance
UKB: Graph Based Word Sense Disambiguation and Similarity

• Knowledge-based approach to word sense classification; no supervision in the form of a manually annotated corpus needed
• Personalized PageRank algorithm
• http://ixa2.si.ehu.es/ukb
First Experiments

• We are using the knowledge graph developed by UKB team via mapping from Bulgarian WordNet to English WordNet

<table>
<thead>
<tr>
<th>Graph</th>
<th>Accuracy</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>WN</td>
<td>0.517</td>
<td>0.940</td>
</tr>
<tr>
<td>WNG</td>
<td>0.538</td>
<td>0.940</td>
</tr>
</tbody>
</table>

• Not very optimistic

• A possible solution: adding more knowledge to the graph
Knowledge Graph

- We performed several extensions of the Knowledge Graph with additional arcs
  - Domain relations from WordNet
  - Inferred hypernymy relations
  - Syntactic relations from the gold corpus
  - Extended syntactic relations
Syntactic Relations

• From Universal Dependency Representation of BulTreeBank extraction of dependency relations denoting event-participant semantic relations: SynSet1 – DepRel – SynSet2

• 15,675 triples

• 8,772 relations: 1,844 nsubj, 3,875 nmod, 1,025 amod, 716 iobj and 1,312 dobj dependency relations
Extended Syntactic Relations

• If in the triple SynSet1 – DepRel – SynSet2, SynSet11 is hyponym of SynSet1 and SynSet1 is participant in the event then we add the triple SynSet11 – DepRel – SynSet2

  A doctor kisses a girl. → A surgeon kisses a girl.

• Result: 372,247 (nsubj), 1,125,823 (nmod), 377,577 (amod), 114,760 (iobj) and 292,202 (dobj) semantic relations
More Syntactic Relations

• The relations in the treebank are not the most general ones

• Our goal for each event to find the most general concept restricting each participant in the event. The same participants in more general event:

  A doctor kisses a girl. → A professional kisses a woman. → A professor kisses a bar girl.

  A doctor kisses a girl. → A doctor touches a girl.

• In the experiments: move to the direct hyperonym and extend with all hyponyms
Knowledge Graph Extensions – Inheritance

professor

professional

doctor

surgeon

touch

kiss

peck
Knowledge Graph Extensions – Syntax
Knowledge Graph Extensions – Syntax
Knowledge Graph Extensions – Syntax
Knowledge Graph Extensions – Syntax V↑

professor

doctor

surgeon

professional

touch

kiss

peck
Knowledge Graphs

- **WN**: WN relations
- **WNG**: WN relations + relations from the glosses
- **WNI**: WN relations + inferred hypernymy relations
- **WNGI**: WN + glosses + hypernymy
- **WNGID1**: WN + glosses + hypernymy + synset-to-domain
- **WNGID2**: WN + glosses + hypernymy + domain synset-to-synset
- **WNGIS**: WN + glosses + hypernymy + dependency relations
- **WNGISE**: WN + glosses + hypernymy + extended dependency
- **WNGISED1**: WN + glosses + hypernymy + extended dependency + synset-to-domain
- **WNGISED2**: WN + glosses + hypernymy + extended dependency + domain synset-to-synset
- **WNGISEUD2**: WN + glosses + hypernymy + extended dependency one level up + domain synset-to-synset
## Results

<table>
<thead>
<tr>
<th>Graph</th>
<th>Accuracy</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>WN</td>
<td>0.517</td>
<td>0.940</td>
</tr>
<tr>
<td>WNG</td>
<td>0.538</td>
<td>0.940</td>
</tr>
<tr>
<td>WNI</td>
<td>0.535</td>
<td>0.940</td>
</tr>
<tr>
<td>WNGI</td>
<td>0.537</td>
<td>0.940</td>
</tr>
<tr>
<td>WNGID1</td>
<td>0.538</td>
<td>0.940</td>
</tr>
<tr>
<td>WNGID2</td>
<td>0.550</td>
<td>0.940</td>
</tr>
<tr>
<td>WNGIS</td>
<td>0.565</td>
<td>0.941</td>
</tr>
<tr>
<td>WNGISE</td>
<td>0.616</td>
<td>0.941</td>
</tr>
<tr>
<td>WNGISED1</td>
<td>0.617</td>
<td>0.941</td>
</tr>
<tr>
<td>WNGISED2</td>
<td>0.624</td>
<td>0.941</td>
</tr>
<tr>
<td>WNGISEUD2</td>
<td>0.656</td>
<td>0.941</td>
</tr>
</tbody>
</table>
Experiments with SemCor

• We have tried the most successful BTB knowledge graph and SemCor as a dataset – non-encouraging result:
  WNG: 57.78; WNGISD2: 57.66; WNGISED2: 55.80

• Syntactic Analysis of SemCor and new semantic relations are extracted: 95901 new relations – new experiments over SemCor itself

• Different domains:
  – The first most frequent synsets in BTB and SemCor do not have common elements
  – About 8000 common relations between BTB and SemCor
OntoNotes

• A corpus with syntactic and semantic annotation
• We are studying the annotation in order to do the same experiments
• Constituent annotation
• Senses are not exactly the same as in WordNet
Knowledge-based WSD – to Sum Up

- Crucial role is played by the Knowledge Graph
- Adding new relations is meaningful and helps
- Searching for new relations – automatic from semistructural information source (efficiency problem)
- Knowledge transfer between languages
- Integration with other approaches
- Integration of annotated texts
Open Questions

• What is a good knowledge graph?
  Hypothesis: similar number of links and disambiguating links

• How to cope with the number of nodes and links?
  Hypothesis: only a small portion of nodes and links converge slowly. Number of iterations is small for many nodes and arcs

• How the nodes from the text are linked to knowledge graph?
  Hypothesis: directed links from KG to the text

• How to incorporate the annotated corpus in KG?
  No idea!
Thanks for your attention.