Using a Wordnet Ontology to Improve the Search of the Digital Dialect Dictionary

Miljana Mladenović, eVox Solutions, Belgrade, Serbia
Ranka Stanković, University of Belgrade, Faculty of Mining and Geology
Cvetana Krstev, University of Belgrade, Faculty of Philology

https://sw4ch2017.ensma.fr/
We will present...

- Method for automatic relating between dialect term and corresponding terms in standard language, [www.vranje.co.rs](http://www.vranje.co.rs)

- The method uses SWRL rules defined in the Serbian WordNet ontology to identify sets of synonymous words.

- It also uses e-dictionaries to produce correct lemmas in the standard language that users usually use for search.

- The method was applied and evaluated on verbs and a group of nouns derived from verbs (verbal nouns).

- We compared results obtained by the system with human evaluators and achieved the accuracy of 89.7%.
Digital dictionary of the South Serbian dialect

- 1st implementation of an on-line dialect vocabulary for Serbian, produced from traditional dialect dictionaries

- \(~20,000\) entries:
  - POS, linguistic information, sound (pronunciation), usage examples, dialect phrases, geolocation, etymology, semantic data, social networks and crowdsourcing.

- Search:
  - by a term, by boolean metadata queries
  - browsing by the 1st letter

http://www.vranje.co.rs
Standard look-up for on-line dictionary.

If user is not familiar with a dialect?

Connecting the standard language and the dialect to enable dialect dictionary search using the standard language terms.
Typical keyword based search

Тща́ша́нъ

-шња, -шњо тадашњи.
„Тъгашњи убави сомуни из фурункъницъ нёма ги више“ (Т.С.).
унео: MZ
Подели реч
Boolean query...

Порекло садржи: ар.
Опис садржи: жена
Порекло садржи: ар.

амамџи́ка ж

жена која ради у амаму (јавном купатилу).
„Прабаба ми је била амамџика кад су били Турци” (Врање).
унео: МЗ
Подели реч

(ар.-тур.) (тур. hamamci)
Semantic search

Претрага  Напредна претрага  Прелистајте ...

по унапред припремљеним кriterијуму:

Пословице  Загонетке  Народне песме  Лична имена  Топоними  Флора

Род  Глаголи  Фигуративни говор  Именице  Вишезначни појмови

крета  аорист  имперфект  глаголска именица  императив

пежоративно  фигуративно  вулгарно  погрдно  деминутив  аугментатив  хипокористик

Укупно нађено 326 записа. Услов претраге: Глаголи - аорист

врнен се

(аор. ја се врна, ти се врна) свр. вратим се.

„Пешки отиде, а на ја се врне” (посл.) (Врање);“
First letter search (filter)

по унапред припремљеним критеријуму:

Пословица  Загонетке  Народне песме  Лична имена  Топоними  Флора

Род  Глаголи  Фигуративни говор  Именице  Вишезначни појмови

Појмови који почињу словом:

А  Б  В  Г  Д Ђ  Е  Ж
З  И  Ј  К  Л Љ  М  Н
Њ  О  П  Р  С  Т Ђ  У
Ф  Х  Ц  Ч  Џ  Ш  Ђ  Đ

Укупно нађено 69 записа. Услов претраге: Đ

Дзђвни

одјекује.
„Удара сас чук, па све дзђвни” (Владовце). „Петлови појев, мори, Морава дзђвни” (нар. пес.) (Врање)
ашлък
мали трошак.
„Дај неки динар, да ми се нађе за ашлък (Биљача).”
унео: MZ
Подели реч

цицијашлък
тврдилук.
„Од њојан цицијашлък по голем га нёма” (Врањска Баньа).
унео: MZ
Подели реч

Биљача, Пчиня District, Serbia
Врањска Баньа, Пчиня District, Serbia
Lexical entry geolocation
Resources for improvement of searching performances

- Serbian morphological e-dictionaries and grammars
  - to produce all inflected forms of standard terms
  - 140,000 lemmas & 5 million forms; 18,000 multi-word lemmas

- Serbian WordNet (SWN) OWL2 ontology
  - rules expressed in Semantic Web Rule Language (SWRL) to generate synonymous groups on the basis of the indirect synonymy relation.

- University of Belgrade
  - Human Language Technology Group
Use of morphological e-dictionaries

- Headword of the verb entry is the present tense, first person singular
- User search for verbs using infinitive
- Infinitive form (lemma) of dialect verb and verb in the standard Serbian (from definition) was added
- After separation of all synonyms aligned with a dialect, infinitive forms were attached to the original form.
- For 3,452 verb entries 7,353 synonyms were detected
  - batalim_**bataliti** | batalen, ostavim_**ostaviti**, napustim_**napustiti**
  - batisujem | kvarim_**kvariti**, upropaštčujem_**upropašćivati**
  - bednim se | lepo se odevam_**odevati**, doterujem_**doterivati** se
  - begam_**begati** | begaj, ja bega_**begati**, ti bega_**begati**, begajeći, bežim_**bežati**
Use of morphological e-dictionaries

- Lemma was assigned for 505 dialect forms out of 3,452 dialect forms given in first person singular, present tense.

- Infinitive forms were assigned to 4,384 word forms in standard Serbian that were connected to dialect forms (out of 7,353).

- Not lemmatized words that consisted of word not presented in e-dictionaries, or adjectives used to describe verbs.

- Relation between verbal nouns and verbs was established in some entries but not systematically.

- In e-dictionaries all verbal nouns are marked with a special marker -> 700 relation were established.
Finding the set of near synonyms by using the WordNet ontology

- Serbian WordNet (SWN), based on Princeton WordNet (PWN) has more than 22,000 concepts (synsets)

- SWN ontology has currently 2,243 verb synsets defined as ontology individuals belonging to the VerbSynset class:
  \[
  \langle \text{rdf:type}} \text{ rdf:resource="&swn30;VerbSynset"/}\rangle
  \]

- Rules: generate synonymous pairs of verbs found in the SWN ontology not based only on the relation of direct synonymy.

- Broader set of synonyms for each verb defined in SWN ontology prodused using relations: synonym, similar to, also see, verb group, hyponym.
Reasoning rules in the SWN ontology

- Eclipse Java EE IDE Luna and Apache Jena for reasoning at the level of OWL 2 language by converting OWL rules into the Jena rules format.

```
"[rule1: (?a eg:label ?b)(?a eg:synonym ?c)(?c eg:label ?e) ->
(?b eg:indirectSynonymy ?e)]"
```

```
"[rule2: (?a eg:label ?b)(?a eg:similar_to ?c)(?c eg:label ?e) ->
(?b eg:indirectSynonymy ?e)]
```

.....

```
"[rule6: (?a eg:similar_to ?c)(?a eg:label ?b)(?c eg:synonym ?d)
(?d eg:label ?e) -> (?b eg:indirectSynonymy ?e)]"
```

- 33 reasoning rules for indirectSynonymy relation
- after inferencing, 6,430 indirectSynonymy related pairs of verbs.
Architecture of the system for building a resource that improves the dialect dictionary search tool

**Extract definitions of verbs in a dialect dictionary, given in standard language**

**Digital Dialect Dictionary**

**Table: dictionary verb entry related with equivalent in standard language**

**E-dictionaries of a standard language morphological transformations for lemma generation**

**Synonym pairs of standard language verbs**

**Jena inferencing tool**

**SWN ontology**

**Index inverting**

**Inverted index table: standard language verb lemma related to equivalent dialect entries**

**Expanded Inverted index table: relation between all standard language verb synonym lemmas and equivalent dialect entries**

**Table: dictionary verb entry related with equivalent standard language lemma of a verb**

**Standard language verb lemma linking to synonyms**
Example

1) Definition extraction
   - isabim
     "(imp. isabi; aor. ja isabi, ti isabi; r.pr. isabija, -ila, -ilo) svr. iskvarim, upropastim."  

2) Lemmatization
   - isabim
     isabi; ja isabi; ti isabi; isabija; iskvarim_iskvariti; upropasti_upropastiti

3) Inverted table
   - upropastiti
     isabim batišem dokrajišem istrovim izabim izakam oznobim profućam

4) Inference rules
   - upropastiti
     unerediti, uništiti, uprskati, zabrljati, zakrmačiti, zasvinjiti

5) Join
   - Upropastiti, unerediti, uništiti, uprskati, zabrljati, zakrmašiti, zasvinjiti
     isabim, batišem, dokrajišem, istrovim, izabim, izakam, oznobim, profućam
Evaluation

- Estimation of the accuracy of pairing the DD and SL entries: 2 language experts annotated the inverted (step 3)
  - Infinitive SL has similar meaning as DD verb?
    - 1 - yes
    - 2 - not clear
    - 3 - no

- Automatic procedure: DD headwords not related to any infinitive

- Infinitive classified ~ take a part in relations 1) related 2) unrelated
  - Human marks 1 with related ⇒ true positives.
  - Human marks 2 and 3 compared to related ⇒ false positives.
  - Comparing with the unrelated set ⇒ false and true negatives.
Evaluation

The confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>System Yes</th>
<th>System No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert yes</td>
<td>$tp = 3022$</td>
<td>$fn = 436$</td>
</tr>
<tr>
<td>Expert no</td>
<td>$fp = 0$</td>
<td>$tn = 784$</td>
</tr>
</tbody>
</table>

- whether dictionary entries are correctly aligned with standard language entries
- P = $tp = (tp + fp) = 1.000$
- R = $tp = (tp + fn)) = 0.874$
- F1 = $2PR = (P + R) = 0.933$
- Accuracy = 0.897

Remarks

- method is completely precise
- FN: shortcomings in the DD
- typos, non-standard verb forms,
- missing SL verb in definition,
- misineterpreted DD verb
Conclusion

Method for improving search of the DD with key-terms in SL

• SL e-dictionaries lemmatize verb forms
• Serbian WordNet based SWRL rules identifies sets of synonymous words for each verb and verbal noun defined in the ontology
• Join two sets of synonym words (from DD and from SL)

Evaluation of the method with data provided by humans

• Accuracy = 89.7%.

Future work

• experiment with other POS
• try to expand the set of ontological rules used in this system