Language technology resources for Estonian text summarization

Kaili Müürisep
University of Tartu
Institute of Computer Science

Outline

- Estonian language resources
- First experiment for summarization
- EstSum ver. 0.1

Institutions

- Institute of Estonian Language: dictionaries, morphology
- Institute of Cybernetics at TTU: text ↔ speech
- University of Tartu: corpus research, morphological analysis, parsing (CG), semantic disambiguation and WordNet, modeling dialogs, research on transcribed speech

Resources

- Corpora: millions of words
- Morphologically disambiguated and syntactically annotated corpora: 300.000 and 100.000 words respectively

Tools

- Morphological analyzer: > 99% get correct analysis, > 45% more than 1analysis
- Constraint Grammar parser:
  - Morphological disambiguator: 85-90% become unambiguos, 2% errors
  - Syntactic analyzer: 83-90% unambiguous, 2-4% errors

AutoSum

- Andres Lippur's bachelor thesis (2000)
- C program
- Extracts relevant sentences
- Uses full syntactic analysis
  - Subjects and objects have extra scores
  - “Black list” of words and POSes
  - Words in title have extra scores
  - Frequency of words
  - First sentences in paragraphs
  - Last sentence of article
- No methodical evaluation

EstSum ver 0.1

- Experiment from 2001 (Institute of Cybernetics)
- Extracts relevant sentences from web version of newspaper Postimees
- Perl program
- Units are words, not sentences
- 3 modules:
  - Html-to-sgml translator
  - Sentence splitter
  - Sentence extractor

HTML-to-SGML

- Webmaster changes his style after every half year
- Ugly html-code
- Impossible to distinguish headers, authors, captions etc.
- This module was also used in another project

HTML-to-SGML example


```
<font face="Verdana, Arial" size="3">
<b>Looduskaitsjad uinutasid küla hirmutanud karu </b></font>
<font face="Verdana, Arial" size="2" color="#CC0000"></font>
<!-- <img hspace=6 vspace=6 border="" src="http://pm.inspiral.net/postimees/failid/kopter.jpg">
 -->
 <br><font face="Verdana, Arial" size="2"><table width=170 align="right"><tr><td><br><font size=1> Alles viies annus uimasteid ja lihaseid halvavat ainet sundis visalt ärkvel püsinud karu alla andma
</td></tr></table>
 <B>Viljandimaa keskkonnateenistuse ja Nigula looduskaitseala töötajatel kulus neljapäeval tunde, et kinni püüda ja Pärnumaa metsasügavustesse viia karu, kes oli päev varem Marna küla elanikke hirmutanud.<P>
 </B><P>
 Marna küla Lohu talu peremees Vahur rääkis, et eelmisel õhtul küla piiranud karu liikus öö
----------------------------------------------------------------------------------------------------------------------------------
<div1 type='unknown'><head>Looduskaitsjad uinutasid küla hirmutanud karu</head>
<p>Pildi allkiri: Alles viies annus uimasteid ja lihaseid halvavat ainet sundis visalt ärkvel püsinud karu alla andma</p>
<p><hi rend='bold'>Viljandimaa keskkonnateenistuse ja Nigula looduskaitseala töötajatel kulus neljapäeval tunde, et kinni püüda ja Pärnumaa metsasügavustesse viia karu, kes oli päev varem Marna küla elanikke hirmutanud.</hi></p>
<p>Marna küla Lohu talu peremees Vahur rääkis, et eelmisel õhtul küla piiranud karu liikus öö
```

Sentence splitter


```
<div1 type='unknown'><head><s>Looduskaitsjad uinutasid küla hirmutanud karu</s></head>
<p>
<s>Pildi allkiri: Alles viies annus uimasteid ja lihaseid halvavat ainet sundis visalt ärkvel püsinud karu alla andma</s>
</p>
<p>
<s><hi rend='bold'>Viljandimaa keskkonnateenistuse ja Nigula looduskaitseala töötajatel kulus neljapäeval tunde, et kinni püüda ja Pärnumaa metsasügavustesse viia karu, kes oli päev varem Marna küla elanikke hirmutanud.</hi></s>
</p>
<p>
<s>Marna küla Lohu talu peremees Vahur rääkis, et eelmisel õhtul küla piiranud karu liikus öö läbi samas ringi ja loom leiti hommikul siitsamast metsast.</s>
</p>
```

Extractor

- $W(s)=\alpha P(s)+\beta F(s)+\gamma K(s)$
- P(s) – position based score of sentence s
  - First sentence of article
  - First sentence of paragraph excluding captions
  - 2^{nd} and 3^{rd} sentence of paragraph
- F(s) – format based score of s
  - Bold gives extra bonus
  - ! and ? in the sentence give penalty score
  - Quatation marks give penalty score
- K(s) – key word score of s (word forms that are frequent in text but not so frequent in general list get extra bonus)

Extractor - example

<div1 type='unknown'><head><s>Looduskaitsjad uinutasid küla hirmutanud karu</s></head>
Min score. 4.14064398
###################

2. p=9.523810 f=12.745098 w=4.957299 s=8.984648 <s><hi rend='bold'>Viljandimaa keskkonnateenistuse ja Nigula looduskaitseala töötajatel kulus neljapäeval tunde, et kinni püüda ja Pärnumaa metsasügavustesse viia karu, kes oli päev varem Marna küla elanikke hirmutanud.</hi></s>

3. p=4.761905 f=4.901961 w=4.628001 s=4.716316 <s>Marna küla Lohu talu peremees Vahur rääkis, et eelmisel õhtul küla piiranud karu liikus öö läbi samas ringi ja loom leiti hommikul siitsamast metsast.</s>

4. p=4.761905 f=4.901961 w=3.012578 s=4.183227 <s>Hommikul üritati teda metsast välja ajada ja tehti tema pihta esimene uinutilask.</s>

Extractor - problems

- No evaluations
- Score values are still preliminary and intuitive
- Word frequencies are calculated on word forms, not roots (14 grammatical cases *2)
- Missing pronoun resolution etc.
- Cohesion problems