Multilingual and cross-lingual news topic tracking

Emilia Käsper[a]

Koke, February 05, 2005

[a] Joint work with the JRC Language Technology Group in Ispra, Italy

Overview

- Geographical place name recognition
  ✓ Geocoding for Estonian
- Hierarchical news clustering
  ✓ News clustering for Estonian
- Cross-lingual news topic tracking

The JRC toolset

- 20 official languages in EU
- **TASK:** Multilingual information retrieval environment
- Lack of linguistic resources
- Lack of experts for maintaining and updating resources
- **SOLUTION:** a linguistically poor solution using mostly statistical tools
- **QUESTION:** can we apply these methods to the Estonian language?

Geocoding: the data

- KNAB database: 22,000 names, 58,000 variants
- ESRI database: 500,000 names
- Geographical information: administrative rank, geographical coordinates
- Locally added: country ISO codes (EE), currency names (Yen), adjectives (British)

Geocoding: the analysis

- Dictionary look-up for capitalised words
- Simple stemming: Sudan's ⇒ Sudan
- Stop-word lists: And (Iran), Split (Croatia), Kerry (USA)
- Multi-word search: New York
- Disambiguation: Paris (FRA) vs 20+ other Parises

Sample HTML output

“Sudanese[As Sūdān/sd] people say goodbye to 20 years of fighting and greet peace,” ran the banner headline in the independent Al-Adhwaa daily. “At last the peace dream has become a reality,” trumpeted its independent rival Al-Rai Al-Aam.

All the papers made much of the rare international spotlight on Sudan [As Sūdān/sd], which saw US [United States of America/us] Secretary of State Colin Powell and other world leaders attend Sundays signing ceremony in Nairobi [Nairobi/ke].

Sample XML information

```
<GEO CID=“SD” PID=“8681” STRING=“Sudan” offset=“629”
DISPNAME=“As Sūdān” DisWeight=“10” CLASS=“0”> Sudan </GEO>

<GEO CID=“US” PID=“719” STRING=“US” offset=“646”
DISPNAME=“United States of America” DisWeight=“10” CLASS=“0”>
US </GEO>

<GEO CID=“KE” PID=“6333” STRING=“Nairobi” offset=“741”
DISPNAME=“Nairobi” DisWeight=“10” LAT=“-1.2702” LON=“36.8041”
CLASS=“1”> Nairobi </GEO>
```

Geocoding of Estonian texts

- Create a local stop-word list
- Morphological preprocessing...?
- Simple stemming makes sense!
  - Sudaanis, Pariisis ⇒ Sudaan, Pariis
  - Itaalias, Veneetsias ⇒ Itaalia, Veneetsia
  - Tallinnas, Kaplinnas ⇒ Tallinn, Kaplinn
  - Yorgis, Frankfurdis ⇒ York, Frankfurt

Geocoding of Estonian texts — problems

- Adjectives in lowercase (briti vs British)
- Systematic misspellings of words with diacritics
  - Šveits ⇒ Shveits, Sveits (Switzerland)
  - Tšehhi ≈ Tshehhi, Tsehhi (Czech Republic)
  - Alžeeria ≈ Alzheeria, Alzeeria, Algeeria (Algeria)

Hierarchical news clustering: the data

- Web crawler visits newsfeeds of news agencies, newspapers, radio stations, tv stations
- Preprocessing removes HTML/XML mark-up, converts to UTF-8
- Word frequency lists for each language
- Global and local stop-word lists

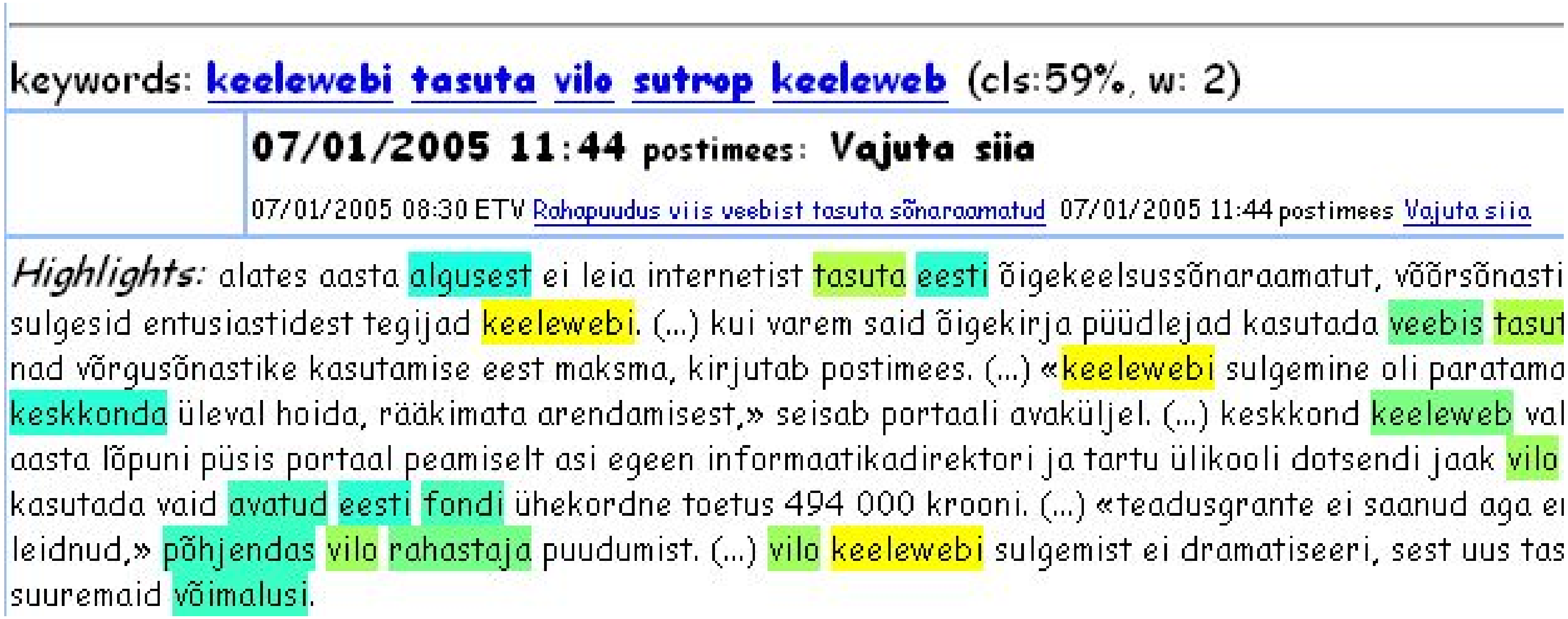
Hierarchical news clustering: the analysis

- Ranked keyword vectors using frequency lists and stop-words
- Ranked country scores from geocoding
- Cosine measure for bottom-up clustering
- Threshold for intra-cluster similarity, no of articles, no of feeds

Clustering of Estonian texts

- Simple stemming not possible
- Full morphological analysis with disambiguation an option
- Local stop-word lists created for both word forms and lemmas
- Gives some results without morphological processing

A sample Estonian cluster



keywords: keelewebi tasuta vilo sutrop keeleweb (cls:59%, w: 2)

07/01/2005 11:44 postimees: Vajuta siia

07/01/2005 08:30 ETV Rahapuudus viis veebist tasuta sõnaraamatud 07/01/2005 11:44 postimees Vajuta siia

Highlights: alates aasta algusest ei leia internetist tasuta eesti õigekeelsussõnaraamatut, võõrsõnasti
sulgesid entusiastidest tegijad keelewebi. (...) kui varem said õigekirja püüdlejad kasutada veebis tasut
nad võrgusõnastike kasutamise eest maksma, kirjutab postimees. (...) «keelewebi sulgemine oli paratama
keskkonda üleval hoida, rääkimata arendamisest,» seisab portaali avaküljel. (...) keskkond keeleweb val
aasta lõpuni püsis portaal peamiselt asi egeen informaatikadirektori ja tartu ülikooli dotsendi jaak vilo
kasutada vaid avatud eesti fondi ühekordne toetus 494 000 krooni. (...) «teadusgrante ei saanud aga ei
leidnud,» põhjendas vilo rahastaja puudumist. (...) vilo keelewebi sulgemist ei dramatiseeri, sest uus tas
suuremaid võimalusi.

Cluster linking across languages: the data

- Eurovoc: a conceptual thesaurus for manual indexing
- Conceptual ⇒ e.g. “protection of minorities”
- Available for 20 languages
- One-to-one descriptor mappings

Cluster linking across languages: the analysis

- Descriptors not explicitly present in text: “protection of minorities” ⇐ “ethnic minority”, “ human right”, “racism”
- Training phase: create associated keyword lists for each descriptor, using a manually indexed test corpus
- Assignment phase: assign descriptors to texts based on keywords
- Map descriptors across languages

Conclusions and future work

- Linguistically poor methods were successfully applied to the Estonian language
- Morphological preprocessing might give further enhancement
- Cross-lingual linking can be employed as soon as Eurovoc becomes available