Experimenting with Slovak Wikipedia as a source for Language Technologies

Michal Laclavík[1], Štefan Dlugolinský[1], Michal Blanárik[2]

[1]Institute of Informatics,
Slovak Academy of Sciences, Bratislava

[2]Faculty of Informatics and Information Technologies, Slovak University of Technology, Bratislava

Wikipedia

- A well-known source of human knowledge maintained by crowd
- A lot of **facts** on various **topics** in a lot of **articles**:
  - 4,405,584 in English
  - 999,839 in Polish
  - 275,982 in Czech
  - 249,421 in Hungarian
  - 187,182 in Slovak
  - 138,292 in Slovene
- http://en.wikipedia.org/wikistats/EN/Sitemap.htm

Wikipedia as a text corpus

- Additional useful information:
  - *articles* represent information about **entities**
  - *links* represent **relations** between **entities**
  - *anchor texts* are **alternative names**, **inflected forms**, **abbreviations** of entities or **entity properties**

- Useful for **NLP tasks** such as NER, QA, MT, WSD, etc.

- NLP Tools based on English Wikipedia:
  - WikipediaMiner
    - detecting and disambiguating Wikipedia topics when they are mentioned in documents
  - Illinois Wikifier
    - disambiguation to Wikipedia with local and global algorithms
  - DBpedia Spotlight
    - tool for annotating mentions of DBpedia resources in text. (Dbpedia is structured information in a from of RDF graphs)

Slovak Wikipedia

- Not explored so far as the English
- Good source of inflected forms and alternative names of entities not included on available dictionaries like Persons, Organizations, Locations

- We made two simple experiments showing possible use of Slovak Wikipedia for NLP:

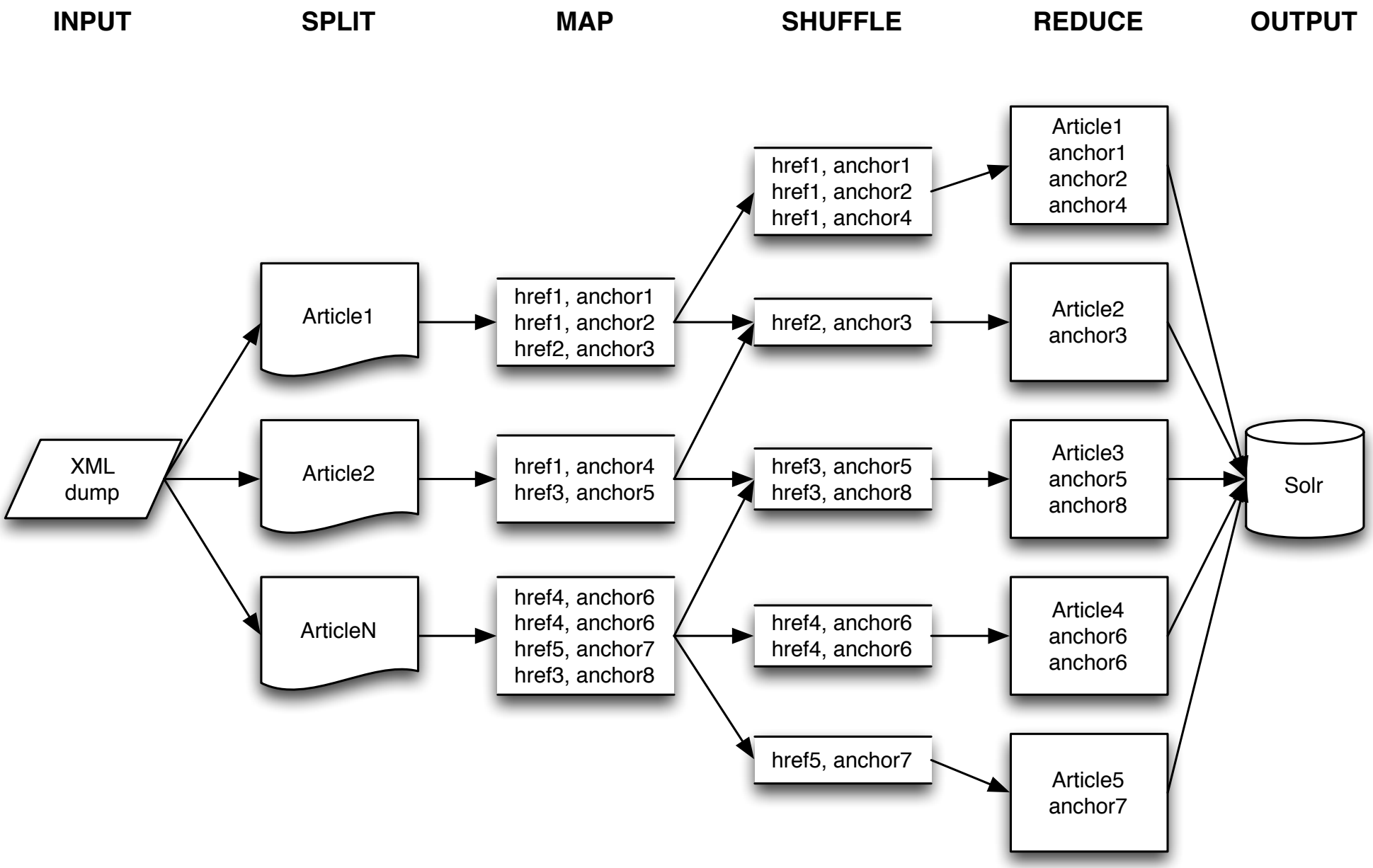
E1: Links and anchors extraction

E2: Named Entity Recognition

E1: Link and anchor text extraction

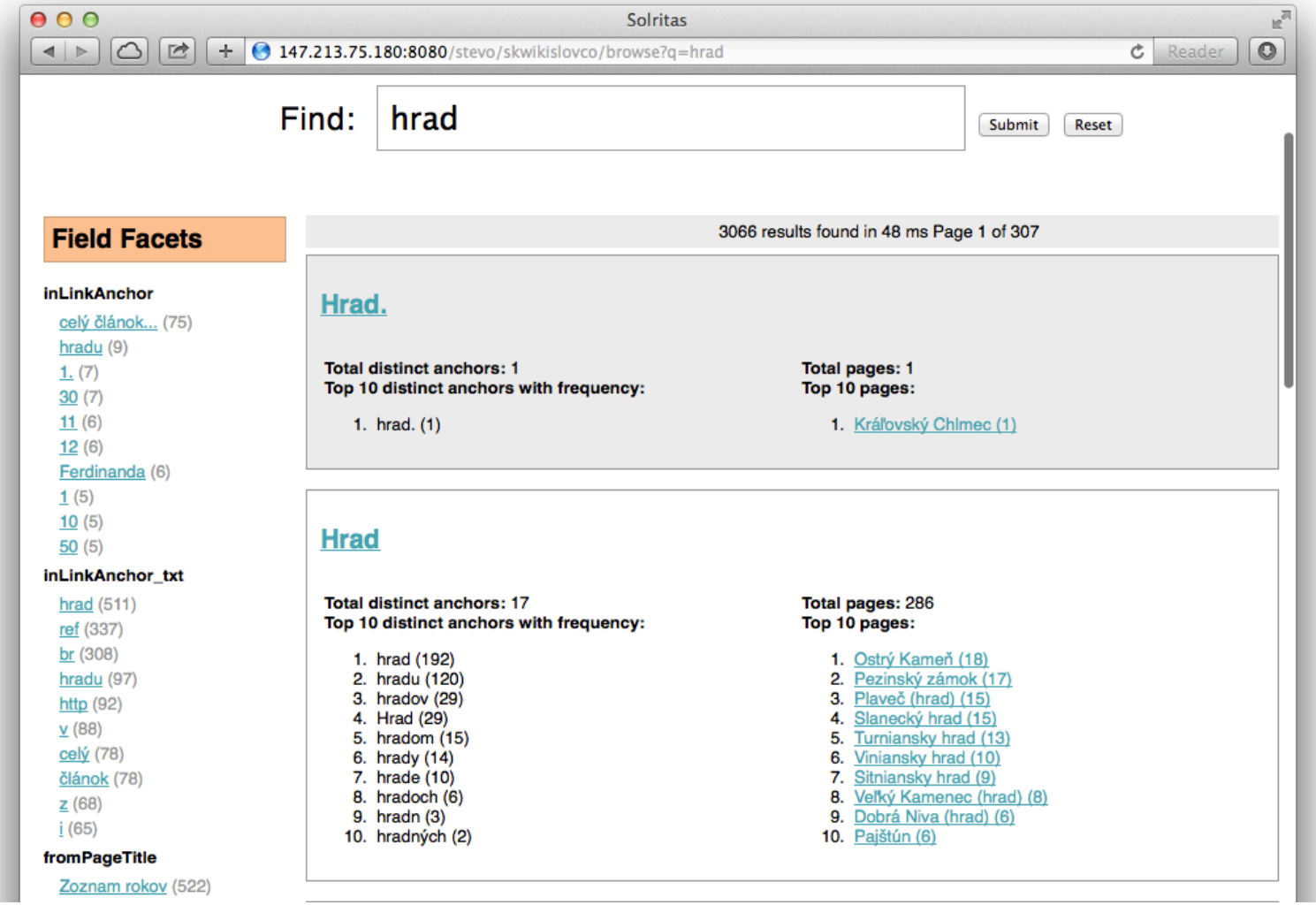
- **The point**
  - Collect entities (articles), their alternative names (anchors) and related entities (via links) and explore search over titles and anchors
- Parsed XML dump of Slovak Wikipedia
  - 737.3 MB uncompressed, 30th April 2013
  - Size of uncompressed English Wikipedia dump is about 44 GB!!!
- We have used **Map-Reduce** paradigm for this task (**Hadoop** implementation)
- Parsed results were indexed in **Solr**

E1: Wikipedia parsing using Map-Reduce



E1 Results

http://147.213.75.180:8080/stevo/skwikislovco/browse?q=hrad



E1 Results

- 310,571 articles processed including redirects from dump XML

- 4,212,467 outlinks with anchor texts extracted
- 3,977,843 inlinks
  - only outlinks to encyclopedia articles, lists, disambiguation pages, and encyclopedia redirects were converted to inlinks

- 696,874 articles indexed including non-existing

- 5.71 inlinks per article in average
- 13.60 inlinks per article (considering only articles referred more than once)

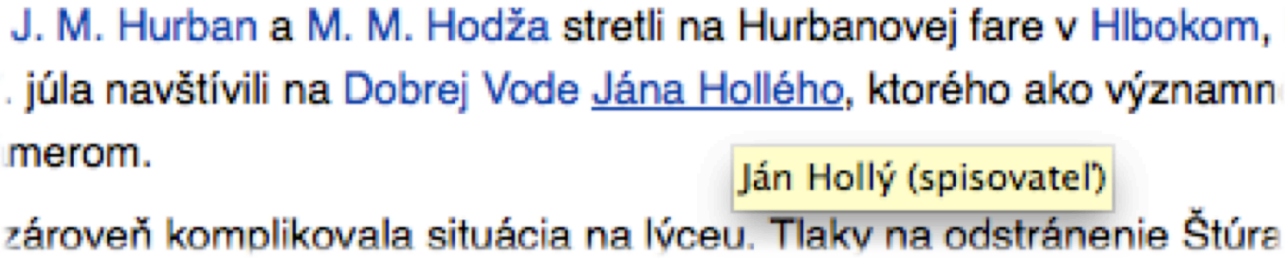
- Wikipedia link structure together with anchor texts could be a resource for creating training sets for NLP methods such as lemmatization, stemming or NER

E2: Named Entity Recognition

- **The point**
  - Annotate persons, locations and their inflected forms in Wikipedia texts and train a NER model on these texts
- Our goal was to automatically train NER model on Wikipedia content and make it applicable on newswire texts for person and location recognition

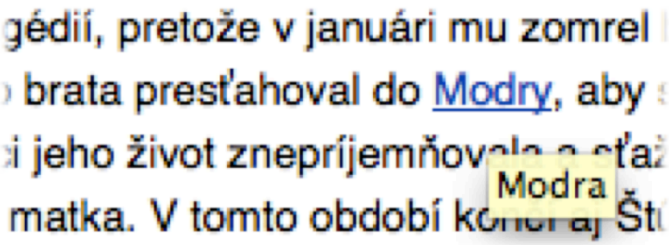
E2: Person names extraction and annotation

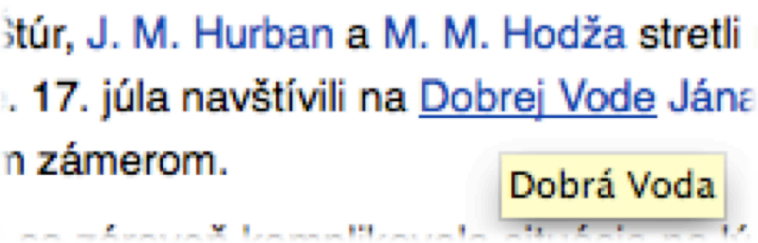
- There were 42,500 unique person names extracted and annotated in Slovak Wikipedia
- Approach:
  - Step 1 (22,511 unique lemmas of person names):
    - 16,454 names in Wikimedia markup; e.g. *[[Ľudovít Štúr]] (* [[1815]])*
    - 11,404 names in Infobox information fields
  - Step 2 (19,989 unique inflected person names):
    - Inflected forms discovered in anchor texts



E2: Location names extraction and annotation

- Location names were extracted and annotated similarly to person names
  - Step 1: 37,121 lemmas of location names
  - Step 2: 482 inflected location names
- Total 37,603 location names





E2: Model training

- Model training

  - Apache OpenNLP used with maximum entropy machine learning (no special tweaking made)
- Trained on annotated Wikipedia texts
  - Example of training data for person NEs:

<START:person> Newtonov <END> interpolačný polynóm alebo presnejšie interpolačný polynóm v **<START:person> Newtonovom <END>** tvare alebo skrátene len **<START:person> Newtonov <END>** polynóm je v numerickej matematike polynóm pomenovaný podľa **<START:person> Isaaca Newtona <END>** interpolujúci danú množinu bodov, ktorý má špecifický tvar, nazývaný '' **<START:person> Newtonov <END>** tvar.

E2: Training data and evaluation

Training data

Sentences	Tagged person names
184,602	91,915

Sentences	Tagged person names
40,579	38,538

Evaluation on Training data

	Precision	Recall	F1
Person model	0,901	0,617	0,867
Location model	0,998	0,995	0,997

Evaluation on Test data – manually annotated news articles

	Precision	Recall	F1
Person model	0,891	0,372	0,517
Location model	1,0	0,292	0,433

E2: Entity merging

- Trained NER models were able to discover entities in text, but not recognize that different forms represent the same entity; e.g.

  [Michael Schumacher], [Michaela Schumachera], [Michaelom Schumacherom]

- Simple merging algorithm based on Levenstein distance and suffixes

E2: Most frequent suffixes of person names found in Wikipedia

Suffix	Frequency	Suffix	Frequency	Suffix	Frequency
a	5 543	ov	130	ea	55
om	4 406	ová	127	s	54
ovi	1 323	ova	103	e	53
m	779	ovho	88	ových	52
ho	589	mu	76	ovom	44
ou	541	ove	71	ému	41
ej	325	ovou	62	o	41
ovej	192	vi	59	ovo	41
ého	189	eho	56	i	40
us	143	ovu	56	*OTHER*	1 345

Conclusion

- Experiments does not provide ready to use solutions for NLP tasks, but show the use pattern of growing Wikipedia resource.

- Our intent was to show that Slovak Wikipedia can serve as a decent source for Language Technology training and evaluation

Thank you!