FACULTY
OF ARTS
Masaryk University

Wikilink:
Its Problems and Possible Solutions

RASLAN 2017, Karlova Studánka, 01 Dec 2017

Vojtěch Mrkývka

Wikilink

- analyzer for Wikipedia articles
- tool for the VisualEditor editation (not a bot)
- looked for new internal links within the article using lemmatized list of names of all articles (the reference list)
- part of my bachelor thesis

Wikilink – preprocessing

- making of the reference list
- no need to recreate it every instance
- time saving

Wikilink – preprocessing

```
[...]
Kostel svaté Barbory
Kostel svatého Augustina
Kostel svatého Michala
[...]
```

Wikilink – preprocessing

```
[...]
Kostel svaté Barbory
Kostel svatého Augustina
Kostel svatého Michala
[...]
```

```
[...]
kostel svatý augustin          Kostel svatého Augustina
kostel svatý barbora           Kostel svaté Barbory
kostel svatý michal            Kostel svatého Michala
[...]
```

Wikilink – link finding

Neverwhere is an urban fantasy television series
by Neil Gaiman that first aired in 1996 on BBC Two

Wikilink – link finding

Neverwhere | is | an | urban | fantasy | television | series
by | Neil | Gaiman | that | first | aired | in | 1996 | on | BBC | Two

Wikilink – link finding

Neverwhere | is | an | urban | fantasy | television | series
by | Neil | Gaiman | that | first | aired | in | 1996 | on | BBC | Two

Neverwhere

Wikilink – link finding

Neverwhere | is | an | urban | fantasy | television | series
by | Neil | Gaiman | that | first | aired | in | 1996 | on | BBC | Two

Neverwhere
IS

Wikilink – link finding

Neverwhere | is | an | urban | fantasy | television | series
by | Neil | Gaiman | that | first | aired | in | 1996 | on | BBC | Two

Neverwhere IS

an

Wikilink – link finding

Neverwhere | is | an | urban | fantasy | television | series
by | Neil | Gaiman | that | first | aired | in | 1996 | on | BBC | Two

Neverwhere	an
IS	by
Urban	that
Fantasy	in
Television	on
...	...

Wikilink – link finding

Neverwhere | is | an | urban | fantasy | television | series
by | Neil | Gaiman | that | first | aired | in | 1996 | on | BBC | Two

Neverwhere	an
IS	by
Urban	that
Fantasy	in
Television	on
. . .	. . .

Wikilink – link finding

Neverwhere | is | an | urban | fantasy | television | series
by | Neil | Gaiman | that | first | aired | in | 1996 | on | BBC | Two

Neverwhere is

Wikilink – link finding

Neverwhere | is | an | urban | fantasy | television | series
by | Neil | Gaiman | that | first | aired | in | 1996 | on | BBC | Two

Neverwhere is
is an

Wikilink – link finding

Neverwhere | is | an | urban | fantasy | television | series
by | Neil | Gaiman | that | first | aired | in | 1996 | on | BBC | Two

Neverwhere is
is an
an urban

Wikilink – link finding

Neverwhere | is | an | urban | fantasy | television | series
by | Neil | Gaiman | that | first | aired | in | 1996 | on | BBC | Two

Urban fantasy
Fantasy television
Television series
Neil Gaiman
First Air

. . .

Neverwhere is
is an
an urban
series by
by Neil

. . .

Wikilink – link finding

Neverwhere | is | an | urban | fantasy | television | series
by | Neil | Gaiman | that | first | aired | in | 1996 | on | BBC | Two

Urban fantasy
Fantasy television
Television series
Neil Gaiman
First Air

. . .

Neverwhere is
is an
an urban
series by
by Neil

. . .

Wikilink – link finding

~~Neverwhere~~ | ~~is~~ | an | urban | fantasy | television | series
by | Neil | Gaiman | that | first | aired | in | 1996 | on | BBC | Two

Wikilink – link finding

Neverwhere | ~~is~~ | ~~an~~ | urban | fantasy | television | series
by | Neil | Gaiman | that | first | aired | in | 1996 | on | BBC | Two

Wikilink – link finding

Neverwhere | is | ~~an~~ | ~~urban~~ | fantasy | television | series
by | Neil | Gaiman | that | first | aired | in | 1996 | on | BBC | Two

Wikilink – link finding

Neverwhere | is | an | urban | fantasy | television | series
by | Neil | Gaiman | that | first | aired | in | 1996 | on | BBC | Two

urban fantasy television
fantasy television series
television series by
Neil Gaiman that
first aired in

Design flaws

- precision of results
- overall speed
- user interface

Design flaws

- precision of results
- overall speed
- user interface

Article	**words**	**time**	**links**	**relevant**
Bergelmir	89	5.45	34	2
Pavel Suchý	420	10.44	95	11
SARS	749	19.53	183	59
Spolek přátel Rumburku	1,171	47.09	319	34
Brno	13,842	393.00	1,978	329

Solutions – precision of results

- use of stop-list on common words like prepositions, conjunctions etc.
- use longer suggestions only
- clustering / already used article-link pairs

Solutions – overall speed

- tree based reference file (for ex. JSON)
- limited use of lemmatisation/desambiguation
- stop-listing and other filtering in preprocessing phase

Solutions – overall speed (JSON example)

```
{
  "kostel": {
    "__ARTICLE__": "Kostel",
    "svatý": {
      "augustin": {
        "__ARTICLE__": "Kostel svatého Augustina"
      },
      "barbora": {
        "__ARTICLE__": "Kostel svaté Barbory"
      },
      "michal": {
        "__ARTICLE__": "Kostel svatého Michala"
      }
    }
  }
}
```

Solutions – user interface

- integration into VisualEditor interface
- use of `Access-Control-Allow-Origin` HTTP header
- partial analysis (one paragraph per time) etc.

Thanks for your attention!

This work was supported by the project of specific research *Čeština v jednotě synchronie a diachronie* (Czech language in unity of synchrony and diachrony; project no. MUNI/A/0915/2016).