PolEval 2020: Information extraction and entity typing from long documents with complex layouts

Filip Graliński

Adam Mickewicz University / Applica.ai

October, 22th

1. INFORMACJE O SPÓŁKACH WCHODZĄCYCH W SKŁAD GRUPY KAPITAŁOWEJ

JEDNOSTKA DOMINUJĄCA – PREZENTACJA SPÓŁKI

Zakłady Magnezytowe „ROPCZYCE" S.A. (ZMR S.A.)

Siedziba: **Ropczyce, woj. podkarpackie**
Adres: **ul. Przemysłowa 1, 39-100 Ropczyce**
Regon: **690026060**
NIP: **818-00-02-127**
www.ropczyce.com.pl

PRZEDMIOT DZIAŁALNOŚCI

Przedmiot działalności ZMR S.A. obejmuje produkcję i sprzedaż zasadowych wyrobów og
które są niezbędnym elementem konstrukcji wyłożeń pieców i urządzeń cieplnych pr
wysokich temperaturach, głównie w hutnictwie żelaza i stali, hutnictwie metali nieżelaz
przemyśle cementowo-wapienniczym, odlewniczym.
Spółka świadczy także usługi w zakresie nawęglania i ulepszania cieplnego wyrobów oraz pr
badawczo-rozwojowe w dziedzinie związanej z przedmiotem jej działalności.

```
period_from   ?
period_to     ?
postal_code   ?
city          ?
...
```

SPRAWOZDANIE Z DZIAŁALNOŚCI GRUPY KAPITAŁOWEJ ZM „ROPCZYCE" S.A. za I półrocze 2012 r.

1. INFORMACJE O SPÓŁKACH WCHODZĄCYCH W SKŁAD GRUPY KAPITAŁOWEJ

JEDNOSTKA DOMINUJĄCA – PREZENTACJA SPÓŁKI

ZM ROPCZYCE S.A.

Zakłady Magnezytowe „ROPCZYCE" S.A. (ZMR S.A.)

Siedziba: **Ropczyce, woj. podkarpackie**
Adres: **ul. Przemysłowa 1, 39-100 Ropczyce**
Regon: **690026060**
NIP: **818-00-02-127**
www.ropczyce.com.pl

PRZEDMIOT DZIAŁALNOŚCI

Przedmiot działalności ZMR S.A. obejmuje produkcję i sprzedaż zasadowych wyrobów og
które są niezbędnym elementem konstrukcji wyłożeń pieców i urządzeń cieplnych pr
wysokich temperaturach, głównie w hutnictwie żelaza i stali, hutnictwie metali nieżelaz
przemyśle cementowo-wapienniczym, odlewniczym.
Spółka świadczy także usługi w zakresie nawęglania i ulepszania cieplnego wyrobów oraz pr
badawczo-rozwojowe w dziedzinie związanej z przedmiotem jej działalności.

`period_from`	2012-01-01
`period_to`	2012-06-30
`postal_code`	39-100
`city`	Ropczyce

...

`company`, `drawing_date`, `period_from`, `period_to`, `postal_code`, `city`, `street`, `street_no`, `people`

Is it just NER?!?

... no!

This is an Information Extraction task, not NER*.

- we are interested in the information not where it is
- not just any person, but CEO, etc.

* But of course you could use NER as a part of the pipeline

Evaluation metric

F1-score will be used as the evaluation metric

It's getting complicated for people

```
[('2012-08-30', 'Józef Siwiec', 'Prezes Zarządu'),
 ('2012-08-30', 'Marian Darłak', 'Wiceprezes Zarządu'),
 ('2012-08-30', 'Robert Duszkiewicz', 'Wiceprezes Zarządu')]
```

Hits/failures will be counted for:

- data-point values: `person__name`, `person__position`, `person__signature_date`,
- relations: `person__name__position`, `person__name__signature_date`

Anna Wróblewska + WUT students: preparing data set
Dawid Lipiński: scripts for converting into Gonito challenge

Any questions related to the challenge?
Filip Graliński, `filipg@amu.edu.pl`

give it a try? `http://poleval2020.nlp.ipipan.waw.pl/challenge/poleval-financial-reports-pl`

quickly get data (without PDFs):

```
git clone git://gonito.net/poleval-financial-reports-pl
```

hands on!

Assume the following ‘process.py‘ Python script

```
#!/usr/bin/python3

import sys
import re


for line in sys.stdin:
    line = line.replace('\\n', ' ')
    m = re.search(r'(Filip|Anna|Janusz) (\S+)', line)
    if m:
        n = m.group(0)
        n = n.replace(' ', '_')
        print('person__name='+n)
    else:
        print('')
```

Let’s evaluate this simple & stupid solution...x

hands on! cntd.

```
git clone git://gonito.net/poleval-financial-reports-pl
cd poleval-financial-reports-pl
wget https://gonito.net/get/bin/geval
chmod u+x geval
xzcat dev-0/in.tsv.xz | cut -f 3 | ./process.py > dev-0/out.tsv
./geval -t dev-0
```

GEval has many more cool features:
https://gitlab.com/filipg/geval#quick-tour

Challenge in numbers

	train	val	test
documents	1628	548	555
size in chars	255.1M	84.2M	90.9M
data points	36519	12314	12700

Table: The data set in numbers

Results

subm.	F_1	P	R
double_big	60.6±1.7	60.8±1.7	60.4±1.8
double_small	58.8±1.8	60.0±1.8	57.7±1.9
middle_big	58.5±1.6	59.9±1.7	57.1±1.7
MBART+RF	44.0±1.4	49.1±1.4	39.8±1.7
CLEX	**65.1±1.9**	**63.8±2.5**	**66.6±1.7**
100_RF	58.4±1.6	56.8±1.6	60.0±1.9
300_xgb	59.2±1.5	57.0±1.6	61.4±1.9
300_RF	58.7±1.5	56.9±1.7	60.5±1.8

Table: The overall results

subm.	pure IE	address	company	date	period	name	posit.
double_big	64.1±1.4	67.1±2.2	59.5±3.9	46.5±4.4	93.8±1.4	69.4±2.4	67.7±2.5
double_small	62.8±1.4	65.9±2.3	58.9±3.6	43.2±4.3	93.4±1.5	67.4±2.5	67.0±2.6
middle_big	62.6±1.4	66.1±2.1	59.6±3.6	42.0±3.9	94.0±1.5	68.2±2.1	65.6±2.6
MBART+RF	49.2±1.3	64.3±2.4	20.5±3.2	19.5±2.6	66.3±3.6	56.9±3.1	59.8±2.9
CLEX	69.7±1.6	81.1±2.1	79.7±3.3	41.0±3.7	97.2±1.2	77.5±2.3	67.8±3.0
100_RF	61.7±1.3	66.6±2.3	59.9±3.9	46.0±4.1	82.7±2.1	67.7±1.9	64.7±2.4
300_xgb	62.8±1.3	67.6±2.1	60.9±4.1	46.3±4.0	90.7±1.7	67.2±2.1	65.4±2.3
300_RF	61.8±1.3	66.7±1.9	57.7±4.1	46.3±4.2	82.1±2.3	68.6±2.1	66.2±2.2

Table: The results for pure information extraction

subm.	all relations	person-position	person-signature
double_big	46.4±2.5	55.4±2.9	**37.4±4.0**
double_small	43.9±2.8	52.8±2.5	34.7±4.3
middle_big	43.0±2.3	52.9±2.7	33.4±3.7
MBART+RF	27.1±2.1	45.5±3.1	8.9±2.4
CLEX	**51.3±2.9**	63.8±2.9	**38.9±4.1**
100_RF	45.2±2.3	52.3±2.5	**38.0±3.6**
300_xgb	44.7±2.2	52.8±2.4	**36.8±3.7**
300_RF	45.9±2.3	53.9±2.7	**37.6±3.6**

Table: The results for relation extraction